



Selected papers from the
CLARIN Annual Conference 2021
Virtual edition



figure
annotations
like
way
selected
form
reference
know
workflow
neural
original
parser
transcrip
de
ty

Selected Papers from the
CLARIN Annual Conference 2021

Virtual Event, 2021, 27–29 September

edited by Monica Monachini 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

189

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

2022

ISBN 978-91-7929-444-1

Introduction

Franciska de Jong

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

Monica Monachini

Programme Committee Chair
Inst. of Computational Linguistics
CNR, Pisa, Italy
monica.monachini@ilc.cnr.it

This volume presents the highlights of the 10th CLARIN Annual Conference 2021. The conference was held on 27th —29th September 2021 and because of the COVID-19 pandemic, for the second year in row a virtual format had to be adopted.

CLARIN, the Common Language Resources and Technology Infrastructure, is a virtual platform that is accessible 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 based on language data in a range of formats and modalities.

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, 2 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 2021 a call² was issued for which 40 abstracts were submitted. The authors of the submissions to the main conference sessions represented 25 countries, both CLARIN ERIC countries (20 members/observers) and countries outside the CLARIN consortia, including Belarus, Brazil, Luxembourg, Spain and Switzerland, to testify the relevance of the CLARIN infrastructure outside the federation, and also externally to Europe.

One of key missions of CLARIN is to foster interactions and synergies between consortia. The Annual Conference represents an excellent opportunity to promote collaboration and this year's event smoothly reflects a good level of cross-country cooperation: out of 35 papers accepted, 10 papers have been written in collaboration by authors of different countries and institutions. The number of the cross-country submissions has been increasing significantly in 2021, passing from 17p.c., registered in 2020, to 29p.c. of papers written in collaboration between different countries. This fact is even more remarkable, if we consider that in 2021 the pandemic dynamics imposed significant mobility restrictions, thus preventing or making collaboration even more difficult.

All submissions were reviewed anonymously by three reviewers (PC members and reviewers invited by PC members). Out of the 40 submitted abstracts 35 submissions were accepted for presentation at the conference (acceptance rate 0.88). The 35 submissions were grouped in the following subjects:

¹<http://www.clarin.eu>

²<https://www.clarin.eu/content/call-abstracts-clarin-annual-conference-2021>

- Annotation and Acquisition Tools (7 papers)
- Legal Issues Related to the Use of LRs in Research (4 papers)
- Repositories and National CLARIN Centres (6 papers)
- Research Cases (3 papers)
- Research Data Management, Metadata and Curation (6 papers)
- Resources (9 papers)

Not surprisingly, Resources, Annotation and Acquisition Tools and Repositories and National CLARIN Centers, each of them representing the kernel of CLARIN, are confirmed as being the areas of major concentration of papers also in this year's Annual Conference, with 9, 7 and 6 papers respectively. Then, Research Data Management and Legal issues, with 10 papers in total, appear to be the other central topics of the Conference. Lastly, 3 papers in total address research questions that require the use of approaches, tools and data available through the CLARIN infrastructure.

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 2018-2020 editions of the CLARIN Annual Conference, a PhD-session was organised with 7 presentations by PhD-students. The abstracts of the student presentations were published in the online programme of CLARIN 2021⁴.

The 2021 edition of the CLARIN Annual Conference was shaped as an online event. The virtual format enabled us to share quality content with almost 425 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 'From Punched Cards to Linguistic Linked Data ...Through Infrastructures'** by Marco Passarotti. The talk discussed how linguistic resources have become increasingly accessible and, lately, interoperable from the very first years of computational linguistics until the present day.
- **Invited talk 'Language Technologies Beyond Research: From Poetry to the Music Industry'** by Elena González-Blanco. The talk showcased the potential of lyrics (the text of songs) analysis for the improvement of recommendation systems in the domain of music, an entertainment in general, in order to achieve better customer experience across different industries.
- **Invited talk 'Language Modeling and Artificial Intelligence'** by Tomáš Mikolov. This talk presented the accomplishments reached so far in statistical language modelling, and scientific challenges that are still in front of us. There is a need to focus more on developing new mathematical models with certain properties, such as the ability to learn continually and without explicit supervision, generalise to novel tasks from limited amounts of data, and the ability to form non-trivial long-term memory.
- **Panel on 'The Role of Corpora for the Study of Language Use and Mental Health Conditions'** was moderated by Henk van den Heuvel with the following experts: Gloria Gagliardi, Stefan Goetze, Saturnino Luz, Khiat Truong.
- During two of the three lunch slots a programme element was offered that allowed additional discussions and networking:
 - **CLARIN Café: "CLARIN Café: Interactive QA Session for Newcomers in CLARIN from the SSH Domain"** with the aim to allow an open question and answering session.

³https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf

⁴<https://www.clarin.eu/content/programme-clarin-annual-conference-2021>

- **Lunch break 'Have Your Lunch with the BoD'** that was open to all participants who were interested to ask direct questions and have a conversation with the members of the Board of Directors.
- **Sessions of accepted conference papers** were organised as regular sessions with a presentation followed by QA.
- During the **CLARIN Student session**, PhD-students presented their work in progress: studies supported by or contributing to the CLARIN infrastructure. The aim of the session was to put the spotlights on the next generation of researchers and enable them to receive feedback on their work from CLARIN experts.
- The **Teaching with CLARIN session**⁵ 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. Three of those submissions were granted with an Award:
 - Teaching with CLARIN Jury Award was awarded to:
 - * Mietta Lennes, Faculty of Humanities, University of Helsinki, Finland, for the 'Introduction to Speech Analysis'⁶
 - * Darja Fišer and Kristina Pahor de Maiti, Faculty of Arts, University of Ljubljana, Slovenia, for 'Voices of the Parliament: A Corpus Approach to Parliamentary Discourse Research'⁷
 - Teaching with CLARIN Audience Award was awarded to Diana Maynard, Faculty of Engineering, University of Sheffield, UK, for 'GATE Training Course'⁸
- As usual the **CLARIN Bazaar** provided an informal setting for conversations with CLARIN people and a space to showcase ongoing work and exchange ideas. The presenters were grouped together by topic in the same breakout rooms to encourage the interaction.
- Each day was finished a **wrap-up session** that combined both personal highlights of two experts in the field and an illustration by professional sketch artist.

In addition, on the event page⁹ 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-2021>
- Recordings of keynote, panel, and CLARIN Café that are available on the CLARIN YouTube channel: <https://www.youtube.com/playlist?list=PLIKmS5dTMgw3MwkJw4fNYsGx994Zi-Gly>.

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 19 (including 1 student paper) full length submissions, out of which 17 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

⁵The slides of this and above mentioned CLARIN Students sessions can be found in the conference programme.

⁶<https://www.clarin.eu/content/introduction-speech-analysis><https://www.clarin.eu/content/introduction-speech-analysis>

⁷<https://www.clarin.eu/content/voices-parliament-corpus-approach-parliamentary-discourse-research>

⁸<https://www.clarin.eu/content/gate-training-course>

⁹<https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event>

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 2021 PC subcommittee were Monica Monachini, Eva Hajičová, Costanza Navarretta, António Branco, Tomaž Erjavec, and Jurgita Vaičenonienė.

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

- Lars Borin, Språkbanken, University of Gothenburg, Sweden
- António Branco, Universidade de Lisboa, Portugal
- Tomaž Erjavec, Jožef Stefan Institute, Slovenia
- Eva Hajičová, Charles University Prague, Czech Republic
- Erhard Hinrichs, University of Tübingen, Germany
- Marinos Ioannides, Cyprus University of Technology (CUT), Cyprus
- Langa Khumalo, South African Centre for Digital Language Resources, South Africa
- Nicolas Larrousse, Huma-Num, France
- Krister Lindén, University of Helsinki, Finland
- **Monica Monachini, Institute of Computational Linguistics “A. Zampolli”, Italy (Chair)**
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
- Costanza Navarretta, University of Copenhagen, Denmark
- 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
- Koenraad De Smedt, University of Bergen, Norway
- Marko Tadič, University of Zagreb, Croatia
- Jurgita Vaičenonienė, 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:

- Olivier Baude, France
- Federico Boschetti, Italy
- Riccardo Del Gratta, ILC “A. Zampolli” CNR Pisa, Italy
- Angelo Mario del Grosso, Italy
- Kinga Jelencsik-Mátyus, Hungary
- Neeme Kahusk, Estonia
- Christophe Parisse, France
- Niccolò Pretto, Italy
- Efstathia Soroli, France
- Thorsten Trippel, University of Tübingen, Germany
- Iulianna van der Lek-Ciudin, CLARIN ERIC, The Netherlands

Contents

Introduction	i
<i>Franciska de Jong and Monica Monachini</i>	
Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant	1
<i>Silvia Calamai, Stefania Scagliola, Fabio Ardolino, Christoph Draxler, Arjan van Hessen and Henk van den Heuvel</i>	
Italian Language Resources. From CLARIN-IT to the VLO and Back: Sketching a Methodology for Monitoring LR's Visibility	10
<i>Dario Del Fante, Francesca Frontini, Monica Monachini and Valeria Quochi</i>	
The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic	23
<i>Isidora Glišić and Anton Karl Ingason</i>	
The TEI-based ISO Standard 'Transcription of spoken language' as an Exchange Format within CLARIN and beyond	34
<i>Hanna Hedeland and Thomas Schmidt</i>	
CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP)	46
<i>Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprunchuk and Valer Varanovich</i>	
Curation Criteria for Multimodal and Multilingual Data: a Mixed Study within the QUEST Project	56
<i>Amy Isard and Elena Arestau</i>	
Legal Issues Related to the Use of Twitter Data in Language Research	68
<i>Pawel Kamocki, Vanessa Hanneschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Lindén and Andrius Puksas</i>	
The Interaction of Personal Data, Intellectual Property and Freedom of Expression in the Context of Language Research	76
<i>Aleksei Kelli, Krister Lindén, Pawel Kamocki, Kadri Vider, Penny Labropoulou, Ramūnas Birštonas, Vadim Mantrov, Vanessa Hanneschläger, Riccardo Del Gratta, Age Värvi, Gaabriel Tavits, Andres Vutt, Esther Hoorn, Jan Hajic Charles and Arvi Tavast</i>	
Collaborating on Language Resource Infrastructures with Non-Research Partners: Practicalities and Challenges	88
<i>Verena Lyding, Egon Stemle and Alexander König</i>	
Annotation Management Tool: A Requirement for Corpus Construction	101
<i>Yousuf Ali Mohammed, Arild Matsson and Elena Volodina</i>	

Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS <i>Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Starkaður Barkarson, Jón Guðnason, Þorsteinn Daði Gunnarsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, Atli Þór Sigurgeirsson, Vésteinn Snæbjarnarson, Steinþór Steingríssson, Gunnar Thor örnólfsson</i>	109
Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction <i>Andrius Utkas, Sigita Rackevičienė, Liudmila Mockienė, Aivaras Rokas, Marius Laurinaitis and Agne Bielinskiene</i>	126
‘Cretan Institutional Inscriptions’ Meets CLARIN-IT <i>Irene Vagionakis, Paola Baroni, Riccardo Del Gratta, Angelo Mario Del Grosso, Federico Boschetti and Tiziana Mancinelli</i>	139
Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts <i>Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Lauriala and Daniela Piipponen</i>	151
Flexible Metadata Schemes for Research Data Repositories. The Common Framework in Dataverse and the CMDI Use Case <i>Jerry de Vries, Vyacheslav Tykhonov, Andrea Scharnhorst, Eko Indarto, Mike Priddy and Femmy Admiraal</i>	168
Bagman – A Tool that Supports Researchers Archiving Their Data <i>Claus Zinn</i>	181
ARCHE Suite: A Flexible Approach to Repository Metadata Management <i>Mateusz Zołttak, Martina Trognitz and Matej Ďurčo</i>	190

Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant

Silvia Calamai
Università di Siena
Siena, Italy
silvia.calamai@unisi.it

Stefania Scagliola
Independent Researcher
Rotterdam, The Netherlands
scagliolas@gmail.com

Fabio Ardolino
Università di Siena
Siena, Italy
fabio.ardolino@unisi.it

Christoph Draxler
Ludwig Maximilian University
Munich, Germany
draxler@phonetik.uni-muenchen.de

Arjan van Hessen
University of Twente
Enschede, The Netherlands
a.j.vanhessen@utwente.nl

Henk van den Heuvel
Radboud University
Nijmegen, The Netherlands
henk.vandenheuvel@ru.nl

Abstract

This paper describes the preparatory phase of a CLARIN-funded project called ‘Voices from Ravensbrück’, which aims to introduce a new type of corpus in the CLARIN resource family called ‘Oral Histories’. The first task consisted in curating and transcribing a set of interviews conducted by the Italian author A.M. Bruzzone with five Italian survivors of the Ravensbrück concentration camp back in 1977. This posed considerable challenges inherent in integrating legacy data from the pre-digital era in the CLARIN infrastructure. The second task was exploring the potential of automatic speech transcription for this type of oral history data. The third element of this exploratory phase was identifying potential partners and suitable data for creating a multilingual collection of existing oral history interviews with survivors of concentration camp Ravensbrück. These preparatory steps were necessary to move to the final phase of our project and realise our overall objective of creating a resource family compliant with CLARIN standards and enabling scholars to analyse interviews from a comparative multilingual and multidisciplinary perspective

1 Italian Interviews: Curation

In 1976, Anna Maria Bruzzone and Lidia Beccaria Rolfi collected testimonies of 5 Italian women who had been deported to the Nazi concentration camp Ravensbrück. The analogue archive containing the Italian interviews was donated to Siena University by Anna Maria Bruzzone’s niece. Siena University digitised all the recordings of Bruzzone’s archive according to IASA standards (.wav format, 96000 Hz, 24 bit). Preservation copies were created and were structured as follows: *i*) audio files, *ii*) photos of the carrier, *iii*) metadata.

Bruzzone’s Ravensbrück collection consists of 14 audio cassettes, with a total duration of about 18 hours and 20 minutes, and contains four long interviews. We know that Anna Maria Bruzzone transcribed the recordings step by step for her publication, writing everything down that she heard. Still, unfortunately, the handwritten transcriptions were lost. Her book, titled *Le donne di Ravensbrück* (Beccaria Rolfi, Bruzzone, 1978, ed. 2021; Figure 1), was divided into 4 sections, each one related to a deportee testimonial (except for the last): Lidia Beccaria Rolfi, Bianca Paganini Mori, Lidia Borsi Rossi, the sisters Nella Baroncini Poli and Lina Baroncini Roveri. In 2016, the book was translated into German (*Als Italienerin in Ravensbrück. Politische Gefangene berichten über ihre Deportation und ihre Haft im Frauen-Konzentrationslager*; Beccaria Rolfi, Bruzzone, 2016). In order to re-use the oral archive and offer insights into valuable legacy data, it was necessary to first define a proper legal framework for

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

Silvia Calamai, Stefania Scagliola, Fabio Ardolino, Christoph Draxler, Arjan van Hessen and Henk van den Heuvel 2022. Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant. *Selected papers from the CLARIN Annual Conference 2021*. Ed. by Monica Monachini and Maria Eskevich. Linköping Electronic Conference Proceedings 189, pp. 1–9. DOI: <https://doi.org/10.3384/9789179294441>



Figure 1: Cover of the volume curated by Lidia Beccaria Rolfi and Anna Maria Bruzzone (2021 edition).

issues related to copyright and privacy. According to Italian law, the copyright of the interviews has been passed on to Bruzzone's niece Paola Chiama and to the heirs of Lidia Beccaria Rolfi, co-author of the volume. In the act of donation to Siena University, P. Chiama has authorised the re-use of the archive for research, dissemination, and teaching purposes (after giving her appropriate advance notice). As for the privacy issues, the legal framework for the re-use and the release (via a CLARIN repository) of the digitized Italian interviews has been accurately defined with the advice of Giuseppe Versaci, lawyer and data protection officer of Siena University and member of the CLARIN Legal and Ethical Issues Committee. In detail, it was considered that, according to Italian law, access to the audio interviews is in any case possible on the basis of the legitimate interest of research forty years after the recordings were made – or sixty years, in case of documents containing data that is likely to reveal the state of health, sexual life or confidential family relationships (Legislative Decree 22/01/2004, art. 122). This normative reference is coupled with art. 2 Legislative Decree 196/2003 (*Italian privacy code*), which gives the heirs the possibility to exercise rights concerning deceased persons (“The rights referred to in Articles 15 to 22 of the Regulation concerning the personal data of deceased persons may be exercised by any entity having a vested interest or acting to protect the data subject as the latter’s agent, or else on household-related grounds deserving protection”).

Notwithstanding this legislative backing, for ethical reasons, it was decided to inform all legal heirs of the five interviewees about the initiated project to obtain their additional consent. With this aim, two distinct documents have been prepared: a) a private letter containing all the information about the project’s backgrounds and goals, and b) a document with detailed information on how the data was processed. The final texts of these documents were submitted to the data protection officer to verify the full compatibility with the Italian and EU legal frameworks.

Once the legal provisions of the project were clear, the heirs of the interviewees had to be traced back. In this phase the support of two Italian associations, the ISR, *Istituto Spezzino per la storia della Resistenza e dell’Età Contemporanea*, and the ANED, *Associazione Nazionale Ex Deportati* was crucial. For all the interviewees, a direct living heir was identified and contacted. Four of the contacted heirs (Aldo Rolfi, Eligio Roveri, Giorgia Poli, and Anna Maria Mori; the last also on behalf of her sister Paola Mori) have notified their full consent to the project, while an answer is still awaited from Borsi Rossi’s heir. Each heir has also received a digital copy of the interview related to his/her relative. Though not set in legal obligations, this approach reflects the type of relationship that oral historians have with their narrators and their next of kin, one in which ‘a shared authority’ on the output of the research is a common practice (Frisch, 1990).

During the project, metadata files compliant with CLARIN standards were created. They were based on an existing CMDI profile developed for an earlier CLARIN interview data curation project named *Oral History Interview* (it can be found as [OralHistoryInterview](#)¹). However, the profile was created for born-digital interview data generated through a new project. Whilst it is true that this profile provides a set of components useful to report pertinent traits of oral interviews (e.g., interviewee and interviewer specifications, interview methods, audio characteristics, and annotation protocols), it only partially adapts to speech materials like the ones contained in the Anna Maria Bruzzone archive: indeed, legacy data pose peculiar challenges on the basis of the relationship among the original documents (analogue carriers), their digitised versions, the documental units (i.e., a single interview which might be contained in more than one single analogue carrier). Therefore, the specificities of Bruzzone’s archive with its original analogue recordings require a partial reorganization of the components of the metadata file, to abide to the archival principle of provenance. The following components must be added: *a*) information about the context in which the interviews were conducted; *b*) information about the process of digitisation of the interviews.

A rearrangement of the original CMDI OH profile component is necessary to meet these new requirements. This could be the addition of two components – *Context of Creation* and *Context of Digitisation* – or the addition of one field with the name *Description* (in the example of the scheme shown in the appendix, the two specific fields have been added). For entering the resource family in the online catalogue of CLARIN, it is important to take into account that three levels of access are required: 1) to the entire collection, 2) to the subcollections in a particular language, and 3) to the single interview. To this end, the [lat-corpus profile](#) of the TLA can be used.

2 Transcription

Under the CLARIN umbrella, starting from the 2016 CLARIN Oral History workshop in Oxford, a group of experts interested in speech data with very different backgrounds – oral history, computational linguistics, anthropology, sociolinguistics, phonetics, and phonology – started exploring how technology can be integrated into research that involves spoken narratives contained in oral archives, thus creating a network called [Speech data and Technology](#)² (Draxler et al., 2020; Scagliola et al., 2020). For making the initial transcriptions, the T-Chain was used. The T-Chain is an open-source transcription workflow for interviews that can be accessed via the [TranscriptionPortal](#)³, where scholars can upload audio files, select the spoken language and, for some languages, a Language Model, and process their files. The digitized version of the five interviews was used as a “stress test” to ascertain the T-Chain’s potential with legacy data. The T-Chain proved to be extremely useful for phonetic and word alignment. At the same time, its use in this preparatory project outlined the workflow’s technical, economic and organizational limits. If one can rely on a previous, accurate transcription, the alignment (both at phonemes and words levels) appears to be rather good. Figure 2 shows a graphical representation of an excerpt from an automatically generated word alignment.

For this alignment, the original speech signal was processed using the Google automatic speech recognizer for the Italian language, which returned an orthographic transcript. This transcript was corrected manually, and then processed with the MAUS web service (Kisler et al., 2012) in order to obtain the alignment. The final result is a three-tiered transcription, which specifies the position of the orthographic words, their expected canonical pronunciations, and the sounds of the actual utterance: the actual utterance is often quite different from the expected pronunciation, due to coarticulation, accented speech, and low signal quality. This position is generally given either in timestamps or sample points relative to the start of the utterance. Table 1 shows the same excerpt as a table, ready for import into a spreadsheet, database system, or statistics package.

¹ <https://catalog.clarin.eu/ds/ComponentRegistry/#/>; look for the profile [OralHistoryInterview](#)

² <https://speechandtech.eu/>

³ <https://speechandtech.eu/oh-portal>

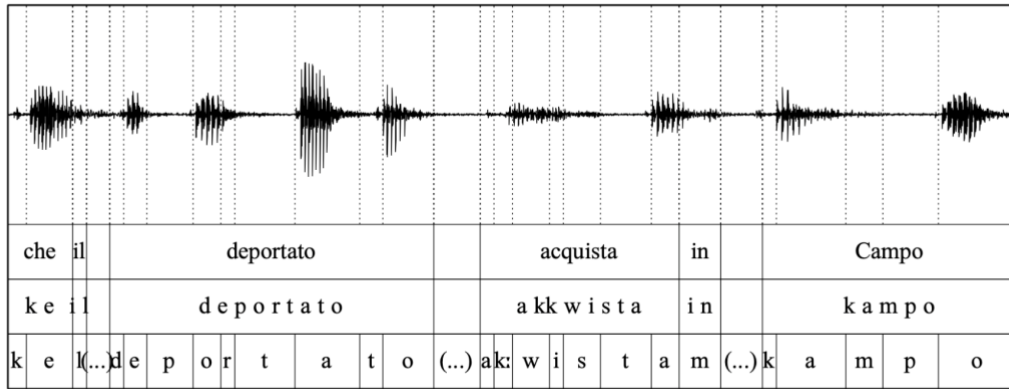


Figure 2: Excerpt of an automatic alignment: the top shows the raw audio signal waveform, the three tiers display the orthographic words, their expected pronunciation, and the actual pronunciation, each with its exact position within the signal.

Line	Begin	Tier	Text
			...
47	21.6800	ORT-MAU	che
48	21.8200	ORT-MAU	il
49	21.8500	ORT-MAU	
50	21.9000	ORT-MAU	deportato
51	22.6000	ORT-MAU	
52	22.7000	ORT-MAU	acquista
53	23.1300	ORT-MAU	in
54	23.2200	ORT-MAU	
55	23.3100	ORT-MAU	Campo
			...

Table 1: Table rendering of the time-aligned orthographic transcription on the word level in Figure 1.

The audio files originally come in archive quality (96 kHz sample rate, 24-bit linear quantisation, stereo) to result in a data rate of 0.58 MB/s. Many speech processing tools do not require such high data rates, and thus the audio files were downsampled to 16 kHz, 16-bit linear quantisation, mono for a data rate of 0.032 MB/s (a >90% reduction). This downsampling is done before the audio data enters the T-Chain to reduce the amount of data to transfer.

After this pre-processing step, we verified the potential of the ASR (*Automatic Speech Recognition*) system for the Ravensbrück interview data. Since we cannot rely on academic open-source ASR software for the Italian language, we were forced to use a commercial system, which obviously was not trained on legacy data. Given the varying acoustic quality of legacy data, the variable interview settings, and the presence/absence of vernacular speech throughout the interviews, we decided to proceed with a qualitative/quantitative evaluation of the effective ASR performances.

The ASR outputs were evaluated for two example audio documents: *i*) the initial part (48 minutes, of which 44.55 spoken) of the interview with Lidia Beccaria Rolfi (file name: BRZTO061a); *ii*) the initial part (31 minutes, of which 29.30 spoken) of the interview with Lina Baroncini Roveri and Nella Baroncini Poli (file name: BRZTO067a).

The two samples have been selected since they represent a pair of opposites: the interview with Lidia appears to be linear, with a sharp prevalence of one of the actors (Lidia), while the interview with Lina and Nella is rather complex, presenting a more intricate diarization among four speakers (Anna Maria Bruzzone, Lidia, Lina, and Nella) and a more vernacular style.

The two ASR outputs for each file have been compared with transcriptions of the same samples that were manually corrected to obtain the highest achievable accuracy. In line with the most widespread standards, a threshold of 30% was fixed for an acceptable Word Error Rate (WER). In both cases, however, the WER significantly exceeded the threshold. The interview with Lidia produced a 37.9% WER, while the one involving Lina and Nella reached a WER of 43% (WER was obtained using the *wersim* package in R). Nevertheless, the adoption of the ASR during the transcription seems to provide valuable help in simplifying the process. The technical limits reported here may be overcome in the (near) future, as well as the economic ones.

On the other hand, legacy data pose huge issues with respect to the quality of the signal because: *a*) in most cases, only a single microphone, placed at a considerable distance from the main speaker, is used, resulting in several speakers in the same audio track, i.e., no channel separation; *b*) practical constraints play a role with regard to respecting audio quality standards during the recording of the interview (unexpected factors); and *c*) the interview style (overlappings, changes in volume, vernacular forms).

In the light of these issues, some changes in the workflow are currently being carried out: i.e., the addition of further chain elements intended for audio definition and noise reduction in order to enhance the performances of the T-Chain. At the same time, the potential of ASR with such data should be emphasized. In case we succeed in collecting a considerable number of interviews from Ravensbrück, we could be able to compute a new language model which will be undoubtedly more effective for this type of legacy data. At present, however, both a manual correction phase for the ASR results and a full-manual transcription for the most complex audio segments remain the only feasible procedure to get the optimal transcription.

In view of the objective to facilitate the use and re-use of the four interviews for various disciplines, the verbatim transcriptions of the digitised audio have been created in such a way as to adhere to the requirements of linguistic research. To this end, the actual dialogues' diarization has been transcribed as it is, even in case of non-linear exchange (e.g., speech overlaps or abnormal turn-takings). Hesitations and reformulations have been transcribed as well, together with other salient acoustic signals (e.g., interjections, laughing). In the transcript, both the discontinuities produced by the actors (e.g., vocalizations like *eeh*, *mmm*, and silences) and the ones produced by contextual factors (e.g., environmental noises, technical issues of the record) have been included. Following such criteria, approximately 1025 minutes have been manually transcribed. The software used to create most of the manual transcript was [OCTRA](#)⁴ (ver. 1.4.3). In some cases, the open-source Audacity software (ver. 3.0.5) was used to increase and decrease the signal amplitude to solve particularly complex speech segments. The transcription was done on an ASUS ZenBook Pro15 laptop running Windows 10.

3 Ravensbrück Multilingual Survey

The second part of the project aimed at reaching out to existing oral history archives and to authors who have used interviews for their publications on Ravensbrück in languages other than Italian, to explore the possibilities for contributing to a multilingual resource family (Calamai et al., 2021). This means that we have also identified some 'uncurated' material that stems, in some cases, from the pre-digital era, such as was the case with the Bruzzone archive. Although the digitisation of this material is beyond the scope of this project and the mission of CLARIN, it is important for the various research communities to know that such material exists. Consequently, in the near future, we will connect to the staff of [EHRI](#)⁵ (European Holocaust Research Infrastructure) to inform them about our findings and endeavours to trace oral history data on Ravensbrück.

In the Netherlands, seven authors of books on Ravensbrück have been traced and contacted through the website of the [Ravensbrück committee](#)⁶, to find material that is not yet published online. In addition,

⁴ <https://clarin.phonetik.uni-muenchen.de/apps/octra/octra/login>

⁵ <https://www.ehri-project.eu/>

⁶ <https://www.ravensbruck.nl/>

the museum of resistance in Amsterdam and the broadcasting company VPRO have also been contacted. With regard to existing online oral history collections, the project can draw on [Getuigenverhalen.nl](https://www.getuigenverhalen.nl/)⁷, which is already directly accessible online. Concerning interviews conducted in English, there is a vast array of online interviews projects from which data could be harvested for the resource family: Shoah Visual History (US), Fortunoff Collection (US), United States Holocaust Memorial Museum (US), Imperial War Museum (GB). The USHMM is the only institute to provide direct online access to metadata, audio/video and transcripts, and has expressed its interest in collaborating and sharing its resources. In other cases, different forms of controlled access have been encountered. With regard to interviews held in German, we have found three online archives to draw on: one [German large-scale video-multilingual archive](https://www.videoarchiv-ravensbrueck.de/de)⁸ initiated by film-maker Loretta Walz, who is interested in collaboration, and two Austrian oral history collections, Erzählte Geschichte and VideoArchiv-Projekt Ravensbrück. To explore the possibility to broaden the range of languages, we have also identified archives in [Poland](https://www.audiohistoria.pl/)⁹ and [Spain](https://ajuntament.barcelona.cat/arxiuunicipal/arxiuhistoric/en)¹⁰.

With regard to the diversity of variables that determine the ease with which the resources can be found and accessed and the level of richness of the data, the following categories can be distinguished:

- a. Analogue or digitised interview data that is not available online but in private hands or at foundations run by volunteers, or held at archives, libraries, and museums, but without direct access because of lack of metadata description (e.g., Bruzzone archive before the CLARIN funded project);
- b. Digitized or digital-born interview data on Ravensbrück that is part of a broader project, that abides to a metadata standard and can be easily identified through a refined search environment and can be directly accessed online (USHMM, Getuigenverhalen);
- c. Same interview data of the point (b), but with access restricted to registered users after the creation of a personal account (Shoah Visual History, Fortunoff);
- d. Digitized or digital-born interview that has been generated for a specific project on Ravensbrück, and are either published online as an autonomous resource (the video archive of Loretta Walz in Germany), or after some time have been integrated into a broader library system. The VideoArchiv-Projekt Ravensbrück can be found in the Austrian mediatheque, but the possibilities for granular searches with regard to the metadata of a specific interview are therefore very limited. The same applies to other projects that have been first created as autonomous entities, but as funding ends, have been integrated into a library system.

Within each category, several variables should be taken into account:

1. The modality of the interview (audio and/or video);
2. The mono- or multi-linguality of the interview;
3. The style (in-depth interviews generated according to rigorous academic standards or more casual interviews filmed at an occasion (international gathering, manifestation, current affairs program);
4. The different categories of survivors involved: political prisoners, resistance fighters, aid workers, Jews, Jehovah's witnesses, gypsies, and groups defined as 'socially deviant' (homosexuals, prostitutes, petty criminals);
5. The relationship between the interviewer and the interviewee.

The richness of perspectives offered by a multilingual resource family about Ravensbrück will offer novel points of view on language diversity within the context of digitisation and public history, conversational styles, and interview styles.

4 Conclusions

This paper shows how the integration of legacy data from the pre-digital era in the CLARIN infrastructure poses considerable challenges. It has been necessary to work on the creation of an *ad-hoc* CMDI profile explicitly devoted to oral history materials originally stored in analogue carrier. At the same time, legacy data can help automatic speech transcription (especially when they appear to be

⁷ <https://www.getuigenverhalen.nl/>

⁸ <https://videoarchiv-ravensbrueck.de/de>

⁹ <https://audiohistoria.pl/>

¹⁰ <https://ajuntament.barcelona.cat/arxiuunicipal/arxiuhistoric/en>

thematically coherent and in a good state of preservation). The project *Voices from Ravensbrück* produced full verbatim transcription of the Italian interviews collected by Anna Maria Bruzzone: a valuable resource for a number of social sciences and humanities sciences, especially in view of its upcoming storage in an accessible CLARIN repository. In parallel, potential partners and suitable data for creating a multilingual collection of existing oral history interviews with survivors of concentration camp Ravensbrück were identified and contacted in order to create a resource family that is compliant with CLARIN standards, enabling scholars to analyse interviews from a comparative multilingual and multidisciplinary perspective. This last aspect, in particular, directly recalls the transnational perspective of the project: at the moment, scholars from five distinct European institutions (specifically from Italy, The Netherlands, and Germany) are involved in the project, each one bringing specific expertise and competencies.

Acknowledgements

The authors wish to thank Lucilla Gigli (Siena University) and Giuseppe Versaci (DPO at Siena University), Marica Setaro (Université de Strasbourg), and the heirs of the deportees (Eligio Roveri, Giorgia Poli, Aldo Rolfi, Anna Maria Mori e Paola Mori).

References

- Beccaria Rolfi, L. and Bruzzone, A. M. 2016. *Als Italienerin in Ravensbrück. Politische Gefangene berichten über ihre Deportation und ihre Haft im Frauen-Konzentrationslager*. Metropol Verlag, Berlin.
- Beccaria Rolfi, L. and Bruzzone, A. M. 2020. *Le donne di Ravensbrück. Testimonianze di deportate politiche italiane*. Einaudi, Turin.
- Calamai, S., Beeken, J., Henk Van Den Heuvel H., Broekhuizen, M., van Hessen A., Draxler C. and Scagliola S. 2021. "Voices from Ravensbrück. Towards the Creation of an Oral and Multilingual Resource Family". In Monachini, M. and Eskevich M. (Eds.), *Proceedings of CLARIN Annual Conference 2021*. Virtual Edition: 16-19. <https://office.clarin.eu/v/CE-2021-1923-CLARIN2021ConferenceProceedings.pdf>.
- Calamai, S., Kolletzek C. and Kelli, A. 2019. "Towards a protocol for the curation and dissemination of vulnerable people archives". In Skadina, I. and Eskevich, M. (Eds.), *Selected papers from the CLARIN Annual Conference 2018*. Linköping University Electronic Press, Linköpings: 28-38. <https://ep.liu.se/ecp/159/003/ecp18159003.pdf>.
- Draxler, C., Van den Heuvel, H., Van Hessen, A., Calamai, S., Corti, L. and Scagliola, S. 2020. "A CLARIN Transcription Portal for Interview Data". In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC2020)*. Virtual Edition: 3346-3352. <http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.411.pdf>.
- Frisch, M. 1990. *A Shared Authority: Essays on the Craft and Meaning of Oral and Public History*. New York University Press, New York.
- Hogervorst, S. 2010. *Onwrikbare Herinnering. Herinneringsculturen van Ravensbrück in Europa*. Uitgeverij Verloren, Hilversum.
- Kisler, T., Schiel, F. and Sloetjes, H. 2012. "Signal processing via web services: the use case WebMAUS". In *Proceedings of Digital Humanities Conference*, Virtual Edition: 30-34. https://pure.mpg.de/rest/items/item_1850150/component/file_1850149/content.
- Scagliola, S., Corti, L., Calamai, S., Karrouche, N., Beeken, J., Van Hessen, A., Draxler, Chr., Van den Heuvel, H., Broekhuizen, M. and Truong, K. 2020. "Cross disciplinary overtures with interview data: Integrating digital practices and tools in the scholarly workflow". In Simov, K. and Eskevich, M. (Eds.), *Selected Papers from the CLARIN Annual Conference 2019*. Linköping University Electronic Press, Linköpings: 126-136. <https://doi.org/10.3384/ecp2020172015>.

Appendix 1

Specimen of metadata profile (Interview with Lidia Beccaria Rolfi)

Component	Component I	Component II	Element	Value
ResourceFamily			<i>Description</i>	Anna Maria Bruzzone Archive - Ravensbrück series
			<i>ID</i>	<i>to be determined</i>
InterviewGeneral			<i>NumberOfSpeaker</i>	2
			<i>Duration</i>	06:56:58
			<i>Owner</i>	Università degli Studi di Siena, Archivio Storico dell'Ospedale Neuropsichiatrico di Arezzo, Arezzo (AR), Italy
			<i>Genre</i>	Interview
			<i>Modality</i>	Audio
	ContextOfCreation		<i>ContextOfCreation</i>	This interview with the Ravensbrück ex-deported Lidia Beccaria Rolfi was originally recorded on a series of analog cassette tapes by the oral historian Anna Maria Bruzzone in preparation of her book (Women of Ravensbruck; original title: Le Donne di Ravensbruck, Einaudi, first publication: 1978). The interview comes from the union of 9 different files, deriving from the digitisation of as many sides of 5 audiocassettes. The original audiocassettes are located in the Archivio Storico dell'Ospedale Neuropsichiatrico di Arezzo (Università degli Studi di Siena), Arezzo (AR), Italy.
		Contact	<i>Role</i>	Coordinator of the project
			<i>Name</i>	Silvia
			<i>Surname</i>	Calamai
			<i>Organisation</i>	Università degli Studi di Siena, Siena, (SI), Italy
			<i>E-mail</i>	silvia.calamai@unisi.it
		Contact	<i>Role</i>	Curator of the analogue collection
			<i>Name</i>	Lucilla
			<i>Surname</i>	Gigli
			<i>Organisation</i>	Università degli Studi di Siena, Siena, (SI), Italy
			<i>E-mail</i>	luccilla.gigli@unisi.it

Component	Component I	Component II	Element	Value
	ContextOfDigitisation		<i>ContextOfDigitisation</i>	The digitisation of the audiocassettes was carried out on 10/01/2019 at the Centro di Sonologia Computazionale (Università di Padova), Padova (PD), Italy.
		Contact	<i>Role</i>	Expert in charge
			<i>Name</i>	Alessandro
			<i>Surname</i>	Russo
			<i>Organisation</i>	Centro di Sonologia Computazionale, Dipartimento di Ingegneria Informatica (Università di Padova), Padova (PD), Italy
			<i>E-mail</i>	alessandro.russo@unipd.it
	DigitalAccess		<i>DigitalAccess</i>	<i>to be determined</i>
			<i>Availability</i>	<i>to be determined</i>
			<i>CatalogueLink</i>	<i>to be determined</i>
	InterviewSummary		<i>InterviewSummary</i>	[0.00 – 15.00] Interview starts in the middle. Lidia says that when she was born, her father was 45 years old. She had a very [0.00 – 15.00] Lidia resumes with her memory of July 25 as an illusion that the war was over, with one brother in military [0.00 – 15.00] Lidia talks about her experience in jail. The first night she was with Carletti and Pina Doleatti, but since they [0.00 – 15.00] Lidia says that in the blocks, there were no toilets. After an interruption, the tapes resumes: Lidia recounts [0.00 – 15.00] The taping resumes from the episode of the Christmas meal. They were punished and then transferred to [0.00 – 15.00] She continues talking about the day of the evacuation. They are told to take everything they need. She takes [0.00 – 15.00] Lidia continues to talk about when they wanted her to get off the train even though she had the same pass [0.00 – 15.00] The following year, she returned to France to meet her companions because she knew the addresses by
	Language		<i>Language</i>	Italian
			<i>Iso-369-3-code</i>	ita
		Multilinguality	<i>Multilinguality</i>	Monolingual
	Interviewee		<i>Name</i>	Lidia
			<i>Surname</i>	Beccaria Rolfi
			<i>BirthPlace</i>	Mondovì (CN), Italy
			<i>BirthCountry</i>	Italy
			<i>ResidentPlace</i>	Mondovì (CN), Italy
			<i>ResidentCountry</i>	Italy
			<i>Role</i>	Former deportee in the Ravensbrück concentration camp, writer, anti-fascist activist
			<i>Family</i>	Parents employed as farmers. Last of six siblings (of which known Rita, Luigi, and Enrico). Mother of one (Aldo Rolfi).
			<i>EthnicGroup</i>	Italian (Piedmontese)
			<i>Age</i>	51 (at the time of the interview). Deceased in 1996.
			<i>BirthYear</i>	1925
			<i>Sex</i>	Female
			<i>Education</i>	Teaching diploma
			<i>Profession</i>	Teacher, writer
			<i>Anonymized</i>	False
		Language	<i>LanguageName</i>	Italian
			<i>Iso-369-3-code</i>	ita
			<i>Description</i>	The spoken variety is the Regional Piedmontese Italian. Local lexical forms are sometimes adopted. Other languages

Component	Component I	Component II	Element	Value
	Interviewer			
			Name	Anna Maria
			Surname	Bruzzo
			Role	Main investigator, historian
			RelationToInterviewee	Shared native city (Mondovi)
			RelationToProject	Author, main investigator
			BirthPlace	Mondovi (CN), Italy
			BirthCountry	Italy
			ResidentPlace	Torino (TO), Italy
			ResidentCountry	Italy
			Family	unknown
			EthnicGroup	Italian
			Age	51 (at the time of the interview). Deceased in 2015.
			BirthYear	1925
			Sex	Female
			Education	Bachelor of Arts, Major in Psychology
			Profession	Teacher, historian, writer
			Anonymized	False
		Language		
			LanguageName	Italian
			Iso-369-3-code	ita
			Description	The spoken variety is the Regional Piedmontese Italian.
InterviewContent				
	InterviewKeywords			
			Keywords	Mondovi; First postwar period; Alpini corps; Fascism; Mussolini; racial laws; Spanish Civil War; Abyssinian War; Second
	Full transcript			
			FullTranscript	Yes
	Coverage			
			SpatialCoverage	Northern Italy, Ravensbrück
			TimeCoverage	1918-1945
InterviewMethod				
			RecruitmentMethod	unknown
			PreInterviewinformatio	unknown
			TypeOfInterview	Free interview
			TopicList	unknown
InterviewAudio				
			AudioFileName	BRZTO061a.wav, BRZTO061b.wav, BRZTO062a.wav, BRZTO062b.wav, BRZTO063a.wav, BRZTO063b.wav, BRZTO064a.wav,
			AudioFormat	.wav
			AudioQuality	Good
			RecordingConditions	Indoors
SpeechTechnicalMetad				
			SamplingFrequency	96 kHz
			NumberOfChannel	2
			ByteOrder	little_endian
			Compression	none
			BitResolution	24 bit
		MimeType		
			MimeType	audio/wav
InterviewAnnotation				
			AnnotationProtocol	time aligned transcript
			CharacterEncoding	ASCII
			AnnotationFileName	BRZTO061a.txt, BRZTO061b.txt, BRZTO062a.txt, BRZTO062b.txt, BRZTO063a.txt, BRZTO063b.txt, BRZTO064a.txt,
			AnnotationType	orthographic, phonetic
			Standards	[Lidia:] speaker
			AnnotationFormat	.txt

Italian Language Resources. From CLARIN-IT to the VLO and Back: Sketching a Methodology for Monitoring LRs Visibility

Dario Del Fante

ILC-CNR - Italy

dario.delfante@ilc.cnr.it

Francesca Frontini

ILC-CNR - Italy & CLARIN ERIC

francesca.frontini@ilc.cnr.it

Monica Monachini

ILC-CNR - Italy

monica.monachini@ilc.cnr.it

Valeria Quochi

ILC-CNR - Italy

valeria.quochi@ilc.cnr.it

Abstract

This paper sketches a user-oriented, qualitative methodology for both (i) monitoring the existence and availability of language resources relevant for a given CLARIN national community and language and (ii) assessing the offering potential of CLARIN, in terms of Language Resources provided to national consortia. From the user perspective, the methodology has been applied to investigate the visibility of language resources available for Italian within the CLARIN central services, in particular the Virtual Language Observatory. As a proof-of-concept, the methodology has been tested on the resources available through the CLARIN-IT data centres, but, ideally, it could be applied by any national data centre aiming to assess the existence of LRs in CLARIN for any given languages and check their accessibility for the interested users. We thus argue that such an assessment might be a useful instrument in the hands of national coordinators and centre managers for (i) bringing to the fore both strengths and critical issues about their data providing community and (ii) for planning targeted actions to improve and increase both visibility and accessibility of their LRs.

1 Introduction

With a distributed network of over 70 centres, CLARIN ERIC's principal aim is to ensure easy access to language resources and tools by researchers from all over Europe and beyond, independently of the original producers, and of the centre or consortium physically hosting them. Therefore, a lot of effort has always been put by CLARIN ERIC into developing and operating central functionalities that would serve this key purpose, to the point that today one does not need to know where a given resource is deposited or even be aware of its existence to be able to find, access and use it. This is achieved also thanks to the CLARIN portal, which acts as a gateway to the whole network's offerings.

The first and foremost central service, the CLARIN virtual *shop window*, is the Virtual Language Observatory, the VLO, (Uytvanck et al., 2010)¹, which makes language resources (LRs) searchable via a unified interface offering faceted search, on the basis of common standardised metadata descriptions. The VLO harvests metadata from all of the official CLARIN data providing centres, as well as from other affiliated catalogues and repositories, e.g. the Europeana catalogue (Eskevich et al., 2017)². Other interesting and useful central discovery services are the Federated Content Search (FCS)³, the Language Resources Switchboard (SB)⁴, and the CLARIN Resource Families⁵. Visibility and usability are directly proportional not only to the quality of the data itself but also, more importantly, of its metadata descriptions. While the central services offer key discovery and data inspection functionalities, because of the

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

¹<https://vlo.clarin.eu>

²See also <https://pro.europeana.eu/post/clarin-and-europeana-make-discovery-and-processing-quick-and-easy-for-135-000-cultural-heritage-objects>

³<https://contentsearch.clarin.eu/>

⁴<https://switchboard.clarin.eu/>

⁵<https://www.clarin.eu/resource-families>

distributed nature of the CLARIN infrastructure, the responsibility of the quality of both data and metadata descriptors ultimately lies within the official repositories and data providing centres of each national consortium and partner.

In this paper we argue that, in order to maximise the visibility of LRs within the CLARIN central services, a good practice for national consortia or national data centres would be to regularly monitor these four “points of access” and analyse how the language resources hosted at their centres or relevant for their research community show up from a user perspective. The activities we propose are complementary to the automated metadata checks that centres can carry out thanks to the CLARIN curation dashboard⁶. What we suggest here is a methodology of a more qualitative nature: an assessment aimed at ensuring that any researcher/end-user can effortlessly find the resources she needs and easily use them as intended. As pointed out by Sugimoto (2016), ‘despite the wide array of useful services for (digital) research in linguistics and the humanities [...] it is unclear whether the community is thoroughly aware of the status-quo of the growing infrastructure’. Such an analysis could thus prove useful for bringing to the fore both strengths and critical issues of their data providing community and for planning targeted actions to improve and increase LRs visibility and accessibility: e.g. (meta-)data curation activities; training events on best practices for data and metadata representation, publication and management; specific outreach and communication activities. For reasons of space, in this work we focus on the VLO and attempt to sketch an analysis of the visibility and searchability of language resources (broadly intended as both data and tools/services) as a useful instrument in the hands of repository managers, consortium managers, user-involvement referents and/or national coordinators for planning recovery or improvement actions, as well as targeted communication and engagement strategies. As a case study, we will look at the resources hosted within the CLARIN-IT consortium as well as resources in or about the Italian language hosted elsewhere, under the assumption that they would be of special interest for the Italian CLARIN user communities.

The paper is organised as follows. Section 2 will briefly present the background and related works we considered. Section 3 outlines the methodology devised as an instrument for monitoring the visibility and searchability of LRs in the VLO. In section 4 we will apply the methodology to CLARIN-IT and to Italian as a use case and proof-of-concept. The insights and take-home messages are discussed in section 5. Finally, the conclusions will briefly anticipate how a monitoring of the other points of access, an issue that will have to be tackled in the near future, might be implemented and how it can be useful.

2 Background Context and Related Works

2.1 The CLARIN-ERIC Four Points of Access

Virtual Language Observatory. The VLO (Uytvanck et al., 2010; Goosen and Eckart, 2014) constitutes the main CLARIN central discovery service, i.e. the principal means of finding and exploring Language Resources, broadly intended as both data and tools/services, which exploits the potential of CMDI metadata (Broeder et al., 2010; Broeder et al., 2012) harvested from all the CLARIN B and C centres and other affiliated repositories. The more data providers make use of shared CMDI metadata profiles for describing their LRs, the more easily findable metadata would be in the VLO, because centrally constant work is dedicated to refining the mapping of metadata onto VLO facets and to improve VLO functionalities. This huge harmonisation effort of course comes at the cost of losing some specificity (that can still be maintained locally) and of a certain degree of fallacy. This is why periodic monitoring and metadata curation also on the side of local repositories would be advantageous.

There are two ways the VLO can be searched: it can be queried, first, with a classic search by keyword terms and the results further filtered using predefined facets. Keyword search also supports a pretty expressive advanced syntax (Goosen and Eckart, 2014) that allows the expert user to perform quite specific searches⁷.

⁶<https://curation.clarin.eu/>

⁷For details see <https://www.clarin.eu/blog/vlo-updated-advanced-search-facilities>

The second one offers faceted browsing: as facets can also be used independently, a user can filter metadata records according to the available categories - the facets - and carry out targeted searches. There are 12 different categories, plus two other useful facets, that can be selected in order to narrow down the selection of displayed records:

1. *Language* - the object language relevant for the resource or tool;
2. *Collection* - the collection to which the resource or tool belongs;
3. *Resource Type* - the type of the language resource (e.g. tool, lexicon, grammar, corpus);
4. *Modality* - the modality of the content of the resource or intended for the tool (e.g. spoken or speech);
5. *Format* - the mime type used in the resource or by the tool;
6. *Keyword* - keywords describing the resource or tool;
7. *Genre* - the genre of the content of the resource (e.g. narrative or conversation);
8. *Subject* - the subject or topic of the content of the resource;
9. *Country* - the country of origin of the source material of the resource;
10. *Organisation* - the organisation currently responsible for the resource or tool, i.e. the holder of distribution rights;
11. *Data provider* - the repository in which the resource is actually deposited and that makes it available;
12. *National project* - the CLARIN national consortium to which the resource pertains.

The Federated Content Search (FCS). Most textual and corpus resources hosted by CLARIN centres can be searched and inspected via dedicated query interfaces or applications run by the centres themselves or by the institutions that own the resources (Odijk, 2017). However, such search interfaces are not always easily usable by first-time users as a vast array of query languages and different implementations can be found, which require time and effort to be learned. In order to spare researchers from learning several new query languages before even being certain that the resource actually meets their needs, CLARIN offers a Federated Content Search (FCS) service⁸, “an integrated search facility to make these unrelated and partly overlapping content search engines available to the research community” from one single point of access and by using a common syntax (Odijk, 2017, p.41). FCS enables a user to enter a single query that is sent simultaneously to multiple search engines at different federated CLARIN centres, which in turn search in the corpora they host. FCS therefore gives users the possibility to conduct a full-text search across all federated resources, or a selection of them. FCS is thus thought of as a first-level exploration of CLARIN corpus- and textual data, which allows the identification of relevant resources.

The Language Resource Switchboard (LRS). Many of the resources that can be discovered through the VLO can be used in many ways. On the one hand, a user can directly download the resources from the local CLARIN repository where these are stored and analyse them offline with her own preferred tools. On the other hand, these resources may be fit to be processed directly by CLARIN tools and services which are available online and distributed over various technical centres in Europe. CLARIN has streamlined this process, thus allowing the users to easily access tools that can be applied to a specific resource by immediately selecting it from the VLO or by using a specific tool: the Language Resource Switchboard (henceforth LRS). LRS therefore is a tool that helps a user find a matching language processing web application for a given resource and process it directly.

⁸<https://www.clarin.eu/content/federated-content-search-clarin-fcs>

The CLARIN Resource Families (CRF, Fišer et al. (2018)) is an initiative aimed at providing a user-friendly overview per data type of the available language resources in the CLARIN infrastructure. The listings for each family are meant to facilitate comparative research and are designed for researchers from the digital humanities, social sciences and human language technologies.

Each family is briefly described and the metadata and the links to the respective download pages and concordancers are displayed. Currently, there are 12 corpora families, 5 families of lexical resources, and 4 tool families.⁹

2.2 Related Works

Considering the variability of the CMDI metadata framework (Haaf et al., 2014) and the fact that the needs of the users may change over time¹⁰, the VLO represents an asset that needs a frequent scrutiny in terms of visibility and searchability of LRs.

Different researchers have approached the analysis of the VLO from different perspectives (among others see Lušický and Wissik (2017) and King et al. (2016)). Particularly, two works have influenced the development of the current methodology. Odijk (2014) approaches the VLO from a critical perspective. He examines the searchability of the resources in the VLO with a special attention to the analysis of the structure of their metadata. He shows that finding data of which it is unknown whether they exist is very difficult and in practice in most cases even impossible, given the widely varying granularity of the metadata descriptors and the fact that metadata are often made in isolation ending up in unnecessary differences. His idea of outlining a method aimed at regularly checking the state of the resources in the VLO has proved to be really important given its dynamic nature and its continuous update with new resources.

On the same line, Odijk (2019) further investigates how to enable an easy discovery of LRs and focuses on tools because the search was not easy and because there were no facets dedicated to software for refining a search. He implements a specific faceted search and proposes a curation procedure to secure the uniformity of descriptors and to make sure that the descriptions based on other profiles correctly contain the relevant information and use the right vocabularies. Odijk fundamentally highlights the necessity for a coordination of a national metadata creation and stresses that every national consortium must reserve economic effort for active participation in the metadata curation task force.

3 Methodology

In this paper, we take into account and complement previous related works by proposing a user-oriented methodology for a qualitative assessment of the visibility and findability of LRs, which can be applicable to every national project. We suggest the following checks should be carried out not only by data managers of new consortia after the registration of at least one B or C centre, but also, periodically, for any national consortium, especially when new centres are registered or new large collections are injected. The aim is to determine the extent to which the resources are correctly and adequately described in terms of metadata descriptors associated to them.

In general, the idea is to explore and test an assessment procedure that may assist repository managers, national coordinators or even the CLARIN central office in harmonising the content of each repository and consequently of the VLO, to the benefit of end-users.

The methodology is composed of two phases. The first phase deals with the inspection of the LRs available from the data centres of a given national consortium, as they show up in the VLO. This is performed by exploiting various facets. The second phase deals with the investigation of the existence in the VLO of the main language(s) of the national project under scrutiny, by means of a systematic analysis of LRs distribution among organisations, collections and data providers outside the national consortium. In what follows we display the structure of this methodology.

⁹<https://www.clarin.eu/resource-families>

¹⁰Hence the urgency to periodically assess also user needs, as done by Monachini et al. (2018), or Lušický and Wissik (2017)

3.1 Phase 1: Check for a National Project

At first, the national consortium of interest should be selected from *the National Project facet*, in order to highlight only the LRs that are provided by its centres. Successively, the results are to be filtered in order to classify the retrieved LRs and check how they are described using the VLP facets, according to four different steps:

1. Languages: check the languages present and calculate the number of LRs for each language;
2. Organisations and Collections - check the Organisations and Collections involved and calculate the number of LRs for each;
3. Resource Type and Data Providers - classify the type of LRs deposited by each data provider;
4. Formats and Availability - Check if:
 - (a) The information on availability is clearly and correctly provided;
 - (b) The items deposited and marked as available have an actionable resource, i.e. a resource that can be downloaded and potentially analysed with offline tools or processed directly e.g. via the LRS.

3.2 Phase 2: Check for a Specific Language

As mentioned in 1, under the assumption that the resources for national language(s) are of particular interest for the community of the corresponding national consortium, we deem it useful to monitor also for the existence, visibility and accessibility of LRs for that specific language(s) outside the consortium centres. This might give national coordinators a useful overview of the presence of their language(s) internationally, and help them plan specific actions in the interest of their communities. In this phase, we thus foresee the following steps:

1. Select the national language(s) from the *Language* tab in order to show only the LRs of interest;
2. Calculate the number of Collections hosting these LRs;
3. Calculate the number of Organisations responsible for these LRs;
4. Calculate the number of Data Providers who deposited these LRs;
5. Compare the results with the information in possession of the national coordination team.

4 Applying the Methodology: the Case of CLARIN-IT and Italian Resources

4.1 CLARIN-IT

At present¹¹ CLARIN-IT has 8 member institutions and two data centres:

- ILC4CLARIN¹², the national CLARIN B data centre, hosted and managed by the CNR Institute for Computational Linguistics “A. Zampolli” in Pisa, the founding member of CLARIN-IT; and
- ERCC¹³, a CLARIN C centre, hosted by the Institute for Applied Linguistics (IAL) at EURAC Research in Bolzano

Both centres offer the whole community to deposit services for the long-term preservation of data in certified repositories that comply with CLARIN requirements and which are regularly harvested by the VLO. The consortium also comprises two K-centres:

¹¹We last updated data on 15 January 2022. As new resources may be constantly deposited, the figures are likely to be different at the time of reading.

¹²<https://ilc4clarin.ilc.cnr.it/>

¹³<https://clarin.eurac.edu/>

- the CLARIN KNOWLEDGE CENTRE FOR DIGITAL AND PUBLIC TEXTUAL SCHOLARSHIP (DIPTeXt-KC)¹⁴, jointly maintained by University Ca' Foscari in Venice and ILC-CNR; and
- the transnational CLARIN KNOWLEDGE CENTRE FOR COMPUTER-MEDIATED COMMUNICATION AND SOCIAL MEDIA CORPORA (CKCMC), a transnational K centre hosted in Italy by the Institute for Applied Linguistics, Eurac Research (IAL) in Bolzano.¹⁵

Currently, CLARIN-IT offers seven different digital collections, which are deposited in one of the two data centres. It thus serves different research sub-communities, particularly oral archives, computer-mediated communication, and digital classics. For the latter it is worth mentioning the effort carried out to facilitate the integration and deposit of important textual collections, such as for instance the *Archivio della Latinità Italiana del Medioevo* (ALIM).

4.2 Phase 1: CLARIN-IT in the VLO - the National Project

Following our methodology, the CLARIN-IT resources in the VLO are easily extracted by filtering the results for *CLARIN-IT* within the national project facet¹⁶, as shown in fig. 1.

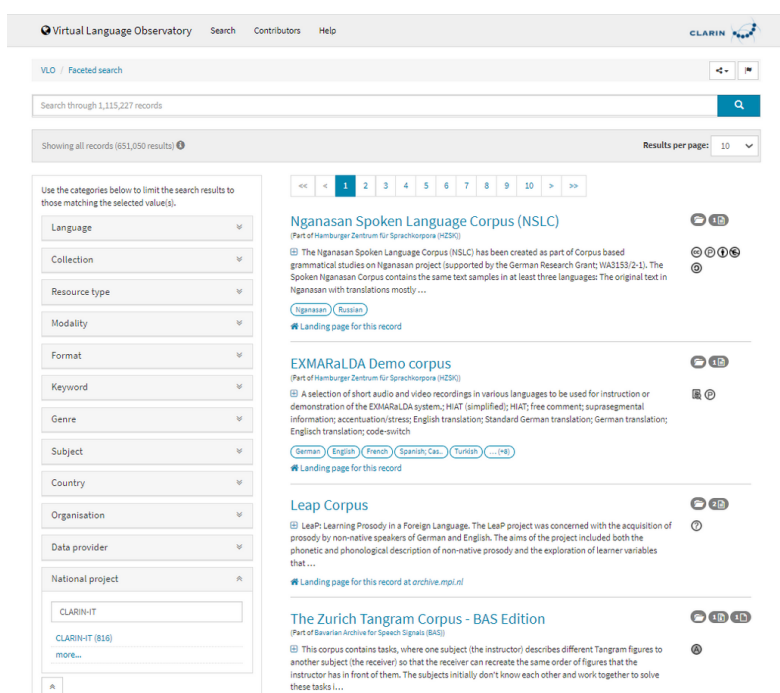


Figure 1: VLO facets - searching for National project.

The query returns 890 different LRs, of which 72 are hidden because they are considered duplicates, which leaves us with 818 distinct resources, as displayed in Table 1.

Duplicate records are automatically hidden in the VLO main search results on the basis of their naming (i.e. title), and are listed under each affected record instead. After a careful examination, most of the apparently duplicate items from our query result in being false duplicates. Within the *ALIM Literary Sources* for instance, all of the 50 hidden items are in fact different critical editions of the same texts made by different editors. Since they have exactly the same title, the system considers them as duplicates. For example, the *Summa Dictaminis* corresponds to three records, one for each editor (Matteo de' Libri, M. Thumser, Emil Polak). While a possible strategy to avoid this could be to add “by EDITOR” to the

¹⁴<https://diptext-kc.ilc4clarin.ilc.cnr.it/>

¹⁵<https://cmc-corpora.org/ckcmc>

¹⁶The executed query is <https://vlo.clarin.eu/search?fqType=nationalProject:or&fq=nationalProject:CLARIN-IT>

<i>Facets</i>	<i>LRs</i>
Languages	27
Organisations	16
Collections	12
Format	10
Resource type	6
Data providers	2
	818 + 72 duplicates

Table 1. CLARIN-IT resources in the VLO along the main dimensions of analysis.

‘title’ (e.g. *Summa Dictaminis BY MATTEO DE’ LIBRI*), this practice would conflict with the decisions taken by the data depositors.

4.2.1 Step-1 - Checking the Languages of the National Project

The first step involves the analysis of the distribution of the languages covered by the resources deposited or described in the CLARIN-IT centres: we identified 27 different languages.

<i>Language</i>	<i>n. LRs</i>	<i>Language</i>	<i>n. LRs</i>
Latin	734	Cimbrian	1
English	44	Croatian	1
Italian	38	Karelian	1
Arabic	32	Ladin	1
German	12	Ladino	1
Ancient Greek	10	Mòcheno	1
Ancient Greek (to 1453)	8	Sardinian	1
Dutch	4	Saurano	1
French	4	Slovenian	1
Czech	2	Spanish; Castilian	1
Modern Greek	2	Trentino	1
Modern Greek (1453-)	2	Tyrolean	1
Basque	1	Veneto	1
Breton	1		

Table 2. Languages in CLARIN-IT.

As Table 2 shows, CLARIN-IT offering is not restricted to Italian only, but it presents LRs in a variety of languages. We acknowledge the substantial presence of LRs in various other languages like English, German, Dutch, French, in addition to a smaller number of LRs in other important European languages like Czech, Modern Greek, Slovenian and Spanish. The wide variety of languages deposited in CLARIN-IT is also evidenced by different regional and minority languages spoken in the North of Italy such as Tyrolean, Trentino, Saurano, Ladin, Cimbrian and Mòcheno¹⁷.

Among the languages, a conspicuous share is represented by Latin and Ancient Greek LRs, and the ILC4CLARIN centre appears to be specialised in hosting them. A closer look, though, reveals an over-representation of Latin LRs, which appear to be more numerous even than the Italian ones. This is due

¹⁷For Sardinian, Karelian, Basque and Breton, there is actually no LR available. In fact, the record refers to a survey run in 2016 by the Digital Language Diversity Project, a dataset containing the original responses to a questionnaire about the online use and usability of these four regional and minority languages.

an organisational choice of the ALIM corpus. As described in Boschetti et al. (2020), every text of that corpus is deliberately deposited as a separate ‘corpus’ resource, so that it reflects the organisation of the original archive; this, however, contrasts with common corpus linguistics practices. Indeed, the peculiar choice of depositing the ALIM corpus as a collection with every document described as a single corpus resource has obvious drawbacks in terms of visibility and searchability, in that it skews the counts in the searches and unnecessarily overloads the search result pages in the VLO. At the same time this structure makes the texts directly and easily actionable by means of NLP services or corpus management tools, available for instance via the Language Resource Switchboard; thus, in terms of accessibility and usability, such an organisation may prove advantageous.

Finally, a similar issue can be observed for Lexical Resources, among which Arabic appears to be over-represented. A closer inspection reveals that the high amount of Arabic LRs - in total 32 - is also due to the way in which the *Al Qamus al Muhit - the Medieval Arabic Lexicon* (Nahli et al., 2016) has been deposited: each letter of the lexicon has been treated as a single deposited resource. Therefore, as in the ALIM archive case, the 30 entries are actually all parts of the same dictionary, i.e. the *Al Qamus al Muhit*.

4.2.2 Step 2 - Checking the Organisations and the Collections Involved

The second step allows us to identify the organisations and consortium members which are actively contributing their resources to CLARIN-IT and analyse how the collections are represented in the VLO for each organisation.

Organisations	Collections	n. LRs	Organisations	Collections	n. LRs
ALIM Archivio della Latinità Italiana del Medioevo	ALIM Literary Sources; ALIM Documentary Sources	344 - 10	Institute of Information Science and Technologies "Alessandro Faedo" ISTI CNR	ILC4CLARIN : ILC Data & Tools	1
DigiLibLT	DigilibLT	364	Escola Universitaria de Turismo "Felipe Moreno" Universitat de les Illes Balears	ERCC Open: CMC & WaC	1
Istituto di Linguistica Computazionale "A. Zampolli" - ILC-CNR	ILC4CLARIN : ILC Data & Tools MQDQ Galaxy	40 - 2	Università del Piemonte Orientale	ILC4CLARIN : OPEN Data & Tools	1
Institute for Applied Linguistics, Eurac Research	Eurac Research: Learner Language; Eurac Research: CMC & WaC	12	CNR Edizioni	ILC4CLARIN : OPEN Data & Tools	1
CIRCSE Research Centre Università Cattolica del Sacro Cuore	CIRCSE	8	Università di Pisa	ILC4CLARIN : OPEN Data & Tools	1
Università di Parma	ILC4CLARIN : OPEN Data & Tools	3	Università di Salerno	ILC4CLARIN : OPEN Data & Tools	1
Ghent University	ERCC: Various; ERCC Open: Learner Language	2 - 1	University of Verona	ERCC: Various	1
Gruppo di ricerca BIA Bibliotheca Iuris Antiqui	BIA-Net FONTES	1	Venice Centre for Digital and Public Humanities (VePDH)	ILC4CLARIN : OPEN Data & Tools	1

Table 3. Organizations and Collections in CLARIN-IT.

In the case of the Italian network, for instance, we identify 16 different organisations currently responsible for 13 collections, as Table 3 shows. Among these, the *Archivio della Latinità Italiana del Medioevo* (ALIM)¹⁸ and the *Digital Library of late antique Latin texts* (DigiLibLT)¹⁹ are responsible for the highest number of Latin LRs. The former is responsible for the *ALIM Literary Sources* and *ALIM Documentary Sources* collections. The latter controls the *DigilibLT* collection. Similarly, Università Cattolica del Sacro Cuore, and more specifically the CIRCSE Research Centre, is responsible for the Latin collection *CIRCSE*, which is composed of Latin lexical resources, corpora and dictionaries, as well as tools for processing Latin texts. ILC-CNR and EURAC, the two CLARIN-IT data providers (cfr. Section 4.1), are directly responsible for 42 and 12 LRs, respectively. ILC-CNR is responsible for two collections: *ILC4CLARIN: ILC Data & Tools*, containing about 40 resources, and the *MQDQ Galaxy* collection, which contains 2 resources²⁰. EURAC is responsible for the *Eurac Research: Learner Language* collection, with 2 resources, and for the *EURAC Research: CMC & WaC* collection, with 4 resources.

As it is clear from Table 3 above, the majority of the organisations are from Italy. Surprisingly, there are, however, two foreign organisations depositing data in Italy: the Ghent University from Belgium, responsible for three annotated learners’ corpora, two in English, French and Dutch, and one in German; and the Universitat de les Illes Balears from Spain, which is responsible for an English lexical resource.

¹⁸<http://en.alim.unisi.it/>

¹⁹<https://digiliblt.uniupo.it/>

²⁰*Musique Deoque* is a project exploring texts of Latin poetry composed in Italy between 1250 and 1550. <http://mizar.unive.it/poetiditalia/public/>

At close examination of the collection *ILC4CLARIN: ILC Data & Tools*, we noticed the presence of additional 17 LRs for which the organisation is not visible from the VLO, although this information is encoded in the full metadata records stored in in the local repository.

This discrepancy may be due to mapping issues between CMDI profiles and VLO facets (already mentioned by Odijk (2019)).

4.2.3 Step 3 - Checking the Resource Types and the Data Providers

As regards the third step, we shall examine how LRs are distributed among the data providers with a focus on *Resource type* in order to get some information on their specialisation. As Table 4 shows, there are two Data Providers in CLARIN-IT: The *ILC4CLARIN Centre* at the Institute for Computational Linguistics and the *EURAC Research CLARIN Centre*.

<i>ILC4CLARIN</i>	<i>n. LRs</i>	<i>Eurac Research</i>	<i>n. LRs</i>
Corpus	375	corpus	18
Text	362	Lexical Resource	1
Lexical Resource	47		
Software, webservice	14		
Language Description	1		
Total	799		19

Table 4. Resource type for each Data provider.

A rich array of LRs is deposited within both of the CLARIN-IT centres. The majority of them are described under the label *corpus* and *text*. The type *lexical resource* corresponds to a broad category and includes lexicons, ontologies, terminologies, e-dictionaries, wordlists etc. . Lastly, the *software* and *webservice* categories include on-line applications, off-line tools and (web)services, which can be used to perform different kinds of analyses on language data.

4.2.4 Step 4 - Checking the Formats and Availability

The last step concerns the query on the available *Formats* and *Subjects* of CLARIN-IT resources. This step allows us to assess whether all resources are correctly deposited and whether they have been further described with suitable and harmonised subject keywords. In the Italian case, the coverage for the latter seems to be incomplete (only 18 LRs are mapped onto VLO subjects keywords, whereas many of the keywords present in the national repositories are not visible in the VLO) and harmonisation could be increased by using controlled vocabularies. One important final check relates to the *Availability* facet, which indicates the “degree to which resources and tools are publicly accessible”. In the case of CLARIN-IT, most of the LRs are publicly available; however, the filter also returned 29 resources with unspecified availability . A closer inspection shows that these correspond to corpora from the ERCC repository and webservices from ILC4CLARIN which are in fact available. This finding might be helpful and lead to amendments of the records.

4.3 Phase 2: Other Resources for Italian in the VLO - an Overview

This second phase aims at investigating the existence, visibility and availability of Italian language resources within the CLARIN ERIC network outside CLARIN-IT ²¹. The idea behind this examination is to shift the perspective from a specific National Project and to focus, instead, on resources of potential interest for the national community, but residing somewhere else.

As indicated in section 3, this phase is composed of 4 steps.

1. Selection of the language. In this use case, we select *Italian* from the language facet and this search gave 9774 LRs present in the VLO.

²¹The query relative to Phase 2 is <https://vlo.clarin.eu/?0&fq=languageCode:code:ita&fqType=languageCode:or>

2. Check for collections: we determined that 94 different collections contain Italian LRs.
3. Check for organisations: there are altogether 109 organisations responsible for the distribution of Italian resources.
4. Lastly, check for data providers: we identified 31 data providers distributed over 15 national projects depositing Italian resources in CLARIN.

The following Table 5 summarises these pieces of information.

Italian in the VLO	
Number of LRs	9774
Collections	94
Organisations	109
Data Provider	31
National Project	15

Table 5. Italian resources in the VLO, provided by organisations outside CLARIN-IT.

By comparing the list of the member organisations of CLARIN-IT with the results obtained from the first test, we end up finding some interesting aspects of the presence of the Italian language in the VLO. First, in addition to Italy and CLARIN-IT, fourteen other countries manage and deposit Italian related LRs. Secondly, and more interesting, there are some institutions located in Italy, not (yet) members of CLARIN-IT, who have deposited some LRs in other CLARIN ERIC centres. For example, the organisation *CNR OVI* appears to be the provider of 5498 resources which are catalogued by *Europeana*²², a European Initiative not directly linked to CLARIN-IT which also catalogues the collections of the *Plutei della Biblioteca Laurenziana*, with 1966 LRs, and the collection of the *Biblioteca Riccardiana*, with 10 LRs. The Italian *Archivio Storico Civico e Biblioteca Trivulziana* located in Milan, which is responsible for the manuscript *Concetti amorosi, cioè lettere giovanili, et amoroze* by Vinetia Compagni, and the *Biblioteca Bertoliana* located in Vicenza, responsible for the *Lettere Amoroze* by Ferrante Pallavicino Luca Assarino, have both deposited these resources to The Language Bank of Finland (FIN-CLARIN). As another example, the University of Naples L'Orientale is responsible for the *MPI EVA corpora: Jakarta Field Station*, a collection of 251 recorded conversations of bilingual Indonesian/Italian children, and deposited it into the *Max Planck Institute for Psycholinguistics* (CLARIAH).

The overview offered by this second test gives us some important insights that can help to enhance the visibility of the Italian language from the perspective of the national consortium, by also highlighting those resources which are not under its management.

5 Lesson Learnt

After having sketched and applied a step-wise methodology for monitoring the availability and visibility in the VLO of LRs, in this section we concentrate on and sum up the lessons we learn from this exercise. While there certainly is ample margin for improvement, we see that the methodology already provides useful insights: on the one hand, it gives useful indications on possible actions that national coordination teams and centre managers may put in place to enhance and promote the consortium offerings; on the other hand, it brings to the fore problematic or controversial issues and thus indicates possible directions for improvements. The results emerging from the different search dimensions may thus help devise targeted communication and engagement actions, or plan recovery strategies. Here below we provide a few examples.

The 'Organisation' dimension in Phase 1 helps us monitor the consortium members' activity as LR providers, and may help coordinators set up actions targeted, for instance, at understanding the reasons

²²<https://www.europeana.eu/en/about-us>

behind a low activity and plan opportune recovery actions, such as specific training events, incentives, focused workshops, or other UI events. At the same time, low activity in providing LR's may instead indicate the given institution has a different profile and might be better involved in other kinds of actions, e.g. in training and UI activities.

The 'Organisation' dimension in Phase 2 helps monitor the activity in depositing data of (national or international) non-member institutions. This might be valuable information for coordinators interested in enlarging their consortium, or in strengthening its international collaborations, as it brings to light the most prominent external providers of LR's of interest for the consortium communities. This may lead to fostering, for instance, mobility grants, joint initiatives, or even new joint K-centres.

The 'Language' dimension in Phase 1 provides interesting indications on the main interests of the active national community and may lead to targeted communication and dissemination actions aimed at maximising national visibility and findability of relevant resources and technology. For instance, national and local websites may be restructured in order to highlight the key resources for the target communities (something similar to the LRF's initiative, but tailored to the national scene). Dedicated info- or training days may be organised at various selected locations, for increasing awareness within the national communities.

The 'Language' dimension in Phase 2 instead provides useful insights on the availability of LR's for the target language(s) from other CLARIN projects and also informs about the interest for one's national language in the rest of the pan-European network. This again may lead to the setting-up of new mobility grants and joint initiatives.

On a more technical, day-by-day operational level, we have seen how the proposed monitoring procedure may also bring to light inconsistencies and problematic issues, for which repository managers may want to plan recovery actions. Some of the identified problematic issues may also be of interest at central level and worth discussing in the appropriate CLARIN ERIC committees, as they may lead to improvements of the central services, as the work by Odijk (2014) already demonstrated. For instance, both the issues of false duplicates and of the granularity of text collections that emerged from Phase 1 (cfr. section 4.2 and 4.2.1) may be an interesting point of discussion both locally and at central level. Locally, it might lead to a confrontation with the responsible of the collection to explore the possibility of aligning their choices to the current majority practices, as well as to targeted data curation activities and metadata refinement.

Centrally, it could raise awareness on peculiar needs of local communities and either lead to an enhancement of central services, or foster the development and dissemination of more stringent and clear common best-practices. Hence, they might be interesting topics for discussion for the Standing Committee for CLARIN Technical Centres and the curation task force.

6 Conclusions

In this paper we have drafted a methodology for monitoring the existence and visibility of LR's of relevance to a national consortium and applied it to CLARIN-IT as a test case. With the growth of the Italian consortium, a thorough check of the LR's contributed by the various actors across the existing official repositories had also become necessary. This qualitative assessment exercise has proven extremely useful and, with adequate extensions and improvements, might become a model for other national projects. Future additions might be to include more dimensions to be assessed, for instance a template search for key resource types, such as reference corpora and lexicons, to ensure that they appear as expected.

While this first assessment focused on the first and most important CLARIN point of access – VLO – similar procedures should be devised also for the other ones. Centres offering NLP web services via the LRS should for example monitor which resources are considered a good match for their tools, and check the outcomes of the analysis, to see if they correspond to the expectations; centres having corpora in the FCS could run test queries to see if selected segments of their corpora actually appear as results in appropriate searches. As for the CRF's, manual curation and updating is currently carried out by a dedicated task force. However, with the growth of the initiative (in terms of families and described items), monitoring of own resources becomes all the more important, for instance to signal changes in resource

size, in the links to online access interfaces, etc. Also, via the CRFs initiative we may discover important resources that are not yet deposited in a CLARIN centre and take steps to encourage their authors to deposit them in a certified repository. Conversely, one may inform the CRFs task force of interesting resources that might be added to the appropriate RF. Thus, monitoring the CRFs and close collaboration with its team can be a source of mutual growth for national repositories and central services, also in terms of reaching out to resource creators that are not yet aware of what CLARIN has to offer them.

Finally, it is important to mention a further, useful central facility offered by CLARIN, which will need to be monitored once it becomes to be more widely used, namely the possibility of creating Virtual Collections²³ both by listing individual resources, and as the outcome of a specific VLO query. Creating a collection such as “Italian historical corpora” could allow users of the Italian national node to access distributed resources at a glance. Such collections, if created from a VLO query, will always be updated with all relevant resources, and in any case provided with persistent identifiers, thus making them easy to be cited in publications.

Acknowledgements

This work has been supported by CLARIN-IT, a Project of International Significance, financed by the Italian MUR, Ministero dell’Università e della Ricerca (Ordinary fund for research institutes).

References

- Boschetti, F., Del Gratta, R., Monachini, M., Buzzoni, M., Monella, P., and Rosselli Del Turco, R. 2020. “Tea for Two”: The Archive of the Italian Latinity of the Middle Ages meets the CLARIN Infrastructure. In Navarretta, C. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference 2020. Virtual Edition*.
- Broeder, D., Kemps-Snijders, M., Uytvanck, D. V., Windhouwer, M., Withers, P., Wittenburg, P., and Zinn, C. 2010. A Data Category Registry- and Component-based Metadata Framework. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- Broeder, D., Windhouwer, M., Uytvanck, D. V., Goosen, T., and Trippel, T. 2012. CMDI: a Component Metadata Infrastructure. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).
- Eskevich, M., Goosen, T., and Uytvanck, D. V. 2017. CLARIN CASE STUDY: Making Europeana’s resources available for research purposes through the CLARIN infrastructure. Technical report, Europeana and CLARIN ERIC. https://pro.europeana.eu/files/Images/Europeana_Research/CLARIN/CLARIN_case_study.pdf.
- Fišer, D., Lenardič, J., and Erjavec, T. 2018. CLARIN’s key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Goosen, T. and Eckart, T. 2014. Virtual Language Observatory 3.0: What’s new? In *Proceedings of the CLARIN annual conference 2014, 23-25 October 2014, Soesterberg, The Netherlands*.
- Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., and Uytvanck, D. V. 2014. CLARIN’s Virtual Language Observatory (VLO) under scrutiny - the VLO taskforce of the CLARIN-D centres. In *CLARIN annual conference 2014*, Soesterberg, The Netherlands.
- King, M., Ostojic, D., Đurčo, M., and Sugimoto, G. 2016. Variability of the Facet values in the VLO – a case for metadata curation. In *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*, pages 25–44. Linköping University Electronic Press, Linköpings universitet.
- Lušický, V. and Wissik, T. 2017. Discovering resources in the VLO: A pilot study with students of translation studies. In Borin, L., editor, *Selected papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 26–28 October 2016*, pages 63–75. Linköping University Electronic Press, Linköpings universitet.

²³<https://collections.clarin.eu/>

- Monachini, M., Nicolosi, A., and Stefanini, A. 2018. Digital classics and CLARIN-IT: What Italian scholars of ancient greek expect from digital resources and technology. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017*, pages 61–74.
- Nahli, O., Frontini, F., Monachini, M., Khan, F., Zarghili, A., and Khalfi, M. 2016. Al qamus al muhit, a medieval Arabic lexicon in LMF. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 943–950, Portorož, Slovenia, May. European Language Resources Association (ELRA).
- Odijk, J. 2014. Discovering resources in CLARIN: Problems and suggestions for solutions. Unpublished paper: <https://dspace.library.uu.nl/handle/1874/303788>.
- Odijk, J. 2017. Introduction to the CLARIN technical infrastructure. In Odijk, J. and van Hessen, A., editors, *CLARIN in the Low Countries*, pages 33–44. Ubiquity Press.
- Odijk, J. 2019. Discovering software resources in CLARIN. In Skadina, I. and Eskevich, M., editors, *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, pages 121–132. Linköping University Electronic Press, Linköpings universitet.
- Sugimoto, G. 2016. Number game - Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. In *Proceedings of the CLARIN Annual Conference 2016, 26-28 September 2016, Aix-en-Provence, France*.
- Uytvanck, D. V., Zinn, C., Broeder, D., Wittenburg, P., and Gardellini, M. 2010. Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, La Valletta, Malta. European Language Resources Association (ELRA).

The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic

Isidora Glišić
University of Iceland
Reykjavik, Iceland
isg14@hi.is

Anton Karl Ingason
University of Iceland
Reykjavik, Iceland
antoni@hi.is

Abstract

The Icelandic L2 Error Corpus is an expanding collection of texts written by users of Icelandic as a second language, published on CLARIN. It currently consisting of 22,705 manually-annotated errors in different categories pertaining to grammar, spelling, lexical and other issues. The corpus was used to perform a contrastive interlanguage analysis, first using a native speaker reference corpus – the Icelandic Error Corpus, then analysing the corpus internally based on linguistic features relevant to second language acquisition. This paper presents the corpus and first results of the analysis.

1 Introduction

Icelandic is a small but increasingly popular language among language learners, both immigrants in Iceland trying to fit into the society and language enthusiasts at large. However, the popularity of Icelandic is a quite novel phenomenon and teaching materials are still scarce and constantly in development. With the rise of language technology efforts in Iceland, it is finally possible to utilize the new technologies in creating ICALL solutions and a major step towards this is creating an error corpus consisting of texts written by users of Icelandic as a second language. At the moment of writing, the Icelandic L2 Error Corpus is a collection of 85 texts, predominantly student essays, annotated for various types of errors. The corpus contains a total of 147,465 words, 15,571 revision spans and 22,705 error instances, where a revision span is a word or a continuous span of words that have been corrected in the annotation process and an error instance is a link between a revision span and a categorization of an error found in the span. This corpus is still likely to grow and is expected to be utilized in analysing learners' interlanguage for the purpose of perfecting teaching materials (both electronic, textbooks and syllabi) and automatic correction tools.

The paper is structured as follows. The theoretical background on learner corpora is discussed in section 2, followed by an overview of previous research on learner interlanguage for Icelandic and the introduction to the new Icelandic L2 corpus in section 2.1. Section 3 describes the methods that we used. Section 4 presents an analysis using two comparative methods.

2 Properties of L2 Mistakes and Error Corpora

The potential that learners' errors have as an indicator of the developmental stages they are likely to have reached in second language acquisition (in further text: SLA) has become obvious already in the 1960s (Thewissen, J., 2013). With the advancement of technology and the rise of corpus linguistics and computer-aided SLA since the early 1990s, more emphasis has been put on the importance of creating learner corpora. These are collections of texts that have been annotated for errors, as this provides access not only to the distribution of learner errors from various perspectives but to their entire interlanguage (Díaz-Negrillo, A., and Fernández-Domínguez, J., 2006), and is key in creating automatic correction tools, development of syllabi, curricula, exams, textbooks and graded readers for SLA.

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

Notable learner corpora include CITE <https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>.

The first step in examining learner corpora is the standardization of error typologies. Most error tagging systems to date tend to be token-based and focus on five distinct linguistic error categories: non-word errors, grammatical errors, lexical errors, errors related to style, and punctuation errors. Non-word errors refer to simple spelling mistakes and accidental repeating of a character or word that result in a word that does not exist, and they are among the most common errors made by native speakers, as well as a specific type of context-related lexical errors which are commonly referred to as confusion sets (Golding, A., and Roth, D., 1999; Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S., 2009; Friðriksdóttir, S. R., and Ingason, A. K., 2020). These are often related to semantically distinct words that are homophones (e.g. *leyti* ‘degree’ and *leiti* ‘hill’ in Icelandic and *piece* and *peace* in English). However, for second language (henceforth, L2) users, grammatical and lexical errors are typically more prominent than in native speakers, and tend to decrease with advancing proficiency level, as the interlanguage is developing closer to the target language. For optimal analysis it is important to observe the size of the corpus and the diversity of submitted texts (both variety of authors and genres). As for the very process of language learning and errors that occur within the interlanguage, there are several factors that need to be considered, some of which are connected to the language situation or task (such as the genre and length of the text and use of reference tools) and others pertaining to the learner (their age, gender, proficiency level, mother tongue and other linguistic background) (Granger, S., 2008).

Scholars have been divided to what extent crosslinguistic influence plays a role. First language (L1) interference was considered crucial in SLA until more extensive research was conducted. Modern theories such as the processability theory emerged, which state that all language learners go through five distinct stages of grammar acquisition, regardless of their native language, and it is not possible to skip a stage or process them in a different order (Pienemann, M., 2011). The theory does not reject crosslinguistic interference but claims that only those linguistic forms that the learner can process can be transferred to the L2 (Pienemann, M., Di Biase, B., Kawaguchi, S., and Håkansson, G., 2005). Therefore, other second languages that the learner acquired before the target language are also relevant, as transference can occur from any other languages that the learner acquired and having internalized more than one grammatical system leads to a generally better understanding of language processing or meta-linguistic awareness (Cummins, J. 1991,). Nevertheless, relevant literature indicates that some classes of common errors are independent of native language background (Gamon, M., Leacock, C., Brockett, C., Dolan, W., Gao, J., Belenko, D., and Klementiev, A., 2013).

An important standard for assessing the stage of learners’ interlanguage is the Common European Framework of Reference for Languages (CEFR) that was launched by the European Council in 2001 as an international standard for describing language ability. It describes language ability on a six-point proficiency scale - A1, A2, B1, B2, C1, C2. A is considered the beginner level, B1 intermediate, B2 advanced and C proficient (near-native) level (Piccardo, E., Goodier, T., and North, B., 2018).¹ The scale is particularly important in evaluating learner errors, as specific types of errors typically emerge on specific proficiency levels. Certain stagnation and regression points have been noted, particularly the one between B1 and B2, where the probability of some types of errors tends to increase rather than decrease. In this case, regression is viewed as a normal part of learning progress, as learners move towards more complex use of language and attempt making longer sentences (Thewissen, J., 2013).

The next section will review previous research on the acquisition of Icelandic as L2 and introduce the L2 error corpus for Icelandic, a novel kind of resource in the context of Icelandic language technology that is still in development.

2.1 Resources for Studying L2 Errors in Icelandic

To understand the context of teaching Icelandic as a second language, one must bear in mind that Icelandic is a small language that was historically spoken by a homogeneous population. For a very long

¹For more details about proficiency level assessment scale, see: <https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>

time, not many foreigners were interested in learning this language and no textbooks or teaching methodology existed. It was not until the 1980s when Svavar Sigmundsson, using contrastive linguistics methods, decided to analyze mistakes of learners of Icelandic, predominantly those of Scandinavian origin (Sigmundsson, S., 1987). As the interest in learning Icelandic as a second language grew, the first textbooks started being published and teaching methodology started developing, setting the standard for the order of grammar acquisition for Icelandic later reiterated and revised many times. However, it was not until very recently that attention was drawn to learner errors in adopting the syllabus to the natural order of acquisition. Using the processability theory, Sigríður Þorvaldsdóttir and María Garðarsdóttir (Þorvaldsdóttir, S., and Garðarsdóttir, M., 2013) started looking into the order of acquisition of cases for their learners' interlanguage on the lowest proficiency levels. Most recently Gísli Hvanndal Ólafsson has examined the general acquisition of grammar from absolute beginners to level A1 (Ólafsson, G. H., 2016) – including cases, verb conjugations and declension of nouns, adjectives and pronouns. Finally the learner corpus was published last year (Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I., 2021) and is at this point still in development. The corpus in its current form is published in a CLARIN repository under a CC BY 4 license.

The Icelandic L2 Error Corpus² currently consists of 85 texts from 36 second language speakers of Icelandic with 15 different first languages, containing 22,705 categorized error instances. Further analysis of the corpus data will follow in section 4. The texts are previously unpublished and obtained directly from their authors, who choose whether the text is to be published under their name or anonymously. The call for texts was first directed to the students of Icelandic as a second language at the University of Iceland but was subsequently extended to a public call. At the moment of writing, the texts are for the most part student essays submitted for evaluation in various courses at the university. The mean number of words per text is 1780 but this number changes drastically when separated by skill level (with the mean for A1 texts being 324 words and 5177 words for C2) as both the written language expression ability and the nature and type of the texts vary - the highest skill level texts typically being long academic essays and parts of or entire MA theses. For more numerical data based on skill level, see Table 4. The currently small size of the corpus (however relatively large when compared with other learner corpora and taking into account the number and variety of the annotated errors) and slow process of acquiring texts are related to the protection of authors' rights, as the University and language school do not have authority to share students' essays and the authors need to submit texts and fill out the publication agreement themselves.

The advantage of using student essays is the accessibility of texts (as it is otherwise very difficult to obtain texts in Icelandic written by foreigners) from subjects with different first and second language background. Furthermore, it is also relatively easy to estimate their proficiency level based on their study progress. The Icelandic as a second language program is separated into a one-year Practical diploma in Icelandic which covers the proficiency level A1-A2 and a 3-year bachelor degree where the students are estimated to be on the level B1-B2 by the end of the first year, and reach B2-C1 by the end of the program (Garðarsdóttir, M., and Þorvaldsdóttir, S., 2020). However, due to the nature of the writings (academic texts) some types of errors tend to be more prominent than in other types of writings. Apart from that, many generic errors might be removed as the texts tend to be polished for better academic success. Texts that arrived from outside of the University were separately scrutinized and the proficiency level was estimated based on the CEFR scale. Bearing in mind the relevant factors in SLA mentioned in section 2, other required information that the subjects provided themselves include their native language, second language(s), length of residence in Iceland and how long they have been learning Icelandic at the time of writing of the submitted text. Other basic demographic information such as age and gender are also part of the form, but are not required. As a relevant number of subjects chose to omit this data, it was not taken into the analysis.

How the corpus was built and the process of extracting and analysing relevant data will be explained in the next section.

²The corpus is available at: <https://repository.clarin.is/repository/xmlui/handle/20.500.12537/106>

3 Methods

The texts for the Icelandic L2 Error Corpus were collected through an open online publication agreement and manually proofread and mapped for errors. Microsoft Word's track changes feature was used for this because it preserves the original version of the text along with the corrected version. After the proofreading process, both versions of the text were extracted and converted, using a Python script, into a single augmented TEI format XML document with labeled enumerated sentences, words and punctuation, and revision spans with unique id numbers containing errors. The errors were analysed and annotated manually and the annotators would label one or several error codes in each revision span. Figure 1 shows an example of a complex revision span containing several error codes and a dependent error.

```
<w>sínum</w>
<revision id="15">
  <original><c>,</c><w>hópurinn</w><w>springur</w><c>,</c><w>og</w><w>svo</w><w>leitt</w></original>
  <corrected><w>sundrast</w><w>hópurinn</w><c>,</c><w>og</w></corrected>
  <errors>
    <error xtype="extra-comma" idx="15-1" eid="0" />
    <error xtype="wording" idx="15-2" eid="0" />
    <error xtype="v3" idx="15-2" eid="0" />
    <error xtype="wording" idx="15-3" eid="0" />
  </errors>
</revision>
<w>annar</w>
<revision id="16">
  <original><w>hlutinn</w><w>af</w><w>sögunni</w></original>
  <corrected><w>hluti</w><w>sögunnar</w><w>leiðir</w></corrected>
  <errors>
    <error xtype="def4ind" idx="16-1" eid="0" />
    <error xtype="wording" idx="16-2" eid="0" />
    <error xtype="dep" depId="15-3" eid="0" />
  </errors>
</revision>
<w>til</w>
<w>harmleiks</w>
<c>.</c>
```

Figure 1: An example of revision spans with multiple error codes and a dependent error.

The figure demonstrates that a revision span can have both multiple codes for different errors, as well as codes which apply to the same error in which case they share the same index (*idx*). So in this example, the error id "15-1" refers to the first character in the revision span, whereas the two words that follow need to be covered by two different error types as there is both invalid syntax (permutation of word order) and a erroneous choice of words involving both of them and both error codes are labeled "15-2". The error code *dep* is not included in the annotation list as its purpose is to annotate that the error in question is a dependent error connected to another in a different revision span, using the original error's *idx*. The corpus syntax along with the error annotation system used for error labeling was originally developed and used for the Icelandic Error Corpus (Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, P., 2020) which contains errors in native speaker texts. However, as the applicability of the corpora extended and they are being used in creating a spelling and grammar correction package for Icelandic,³ it became evident that relying on this formatting of revisions spans is sub-optimal, as it makes it impossible to know automatically which subset of the tokens of the revision spans are concerned with which errors. The spans are therefore being revised in the new version of the corpora and should ultimately include information that should connect each word in the span to a specific error it is related with.

The annotation system that was originally created has also undergone changes in the process of creating the L2 corpus, as new labels needed to be added for errors that were specific for second language use (the list of errors specific to the L2 corpus can be viewed in Table 2). The error tagset consists of 6

³The correction package, *Greynir Correct*, is available at: <https://github.com/mideind/GreynirCorrect>

main categories (*coherence, grammar, orthography, style, vocabulary, other*) which are further divided into subcategories. Some subcategories are very narrow while others are more wide-ranging (notably the Orthography-Punctuation category) and in total there are 259 error codes. The codes are meant to be short but descriptive and are often abbreviated versions of the issue they pertain to, e.g. *simple4cont* stands for simple-for-continuous, where a verb is used in the simple tense and should be in the continuous tense; *agreement-pred* signifies that a predicate is not in agreement with its subject. Some error codes are more specific, such as *geta*, which indicates that the Icelandic auxiliary verb ‘geta’ (e.’be able to’) is wrongly used with an infinitive or present tense instead of past participle. A list of all the codes along with an example and a description is available at <https://github.com/antonkarl/iceErrorCorpusSpecialized/blob/master/errorCodes.tsv>.

After the dataset of TEI documents has been finalized (note that new texts are still likely to be added and the corpus is a work in progress), statistical analyses were conducted that included quantifying the number of texts, revision spans and error occurrences in the corpus, contrasting the L2 error corpus with the Icelandic Error Corpus by ranking the frequency of the error codes extracted as the number of errors per 1000 words. Moreover, each document contains metadata including the author’s first language, other languages, length of residence in Iceland, length of study of Icelandic, and proficiency level. This data is stored to extract specific information on errors based on these parameters and will be analysed in the next section.

4 Data Analysis

As stated before, learner corpora can provide invaluable insight into the learners’ interlanguage, uncovering various linguistic features depending on the variables that the analysis focuses on. The method primarily used for this purpose is contrastive interlanguage analysis (CIA) which compares varieties within one language using two types of comparison: comparing learner language with native speaker reference corpora (L2 vs. L1) or comparing different varieties of learner language (L2 vs. L2) (Granger, S., 2008). The former can uncover the distinguishing features of L2 language use while the second allows us to assess the generalizability of interlanguage features across different factors, learner and task based. As an error corpus for L1 Icelandic has recently been finalized, this provides us with the possibility to make a CIA based on the first type mentioned, and the results will be presented in the following section. For the L2 vs. L2 analysis, as the corpus is still small and the distribution of features such as age or mother tongue is not as wide, the focus will be on the proficiency level and length of residence, which tend to intertwine.

4.1 General Characteristics of L2 Errors Comparative to L1 Errors

Corpus	Files	Total words	Revisions	Categorized Errors	Errors/1000w
Icelandic Error Corpus	4,046	1,137,941	44,261	55,346	44.56
Icelandic L2 Error Corpus	85	147,465	15,571	22,705	153.97

Table 1: Numerical data for both L1 and L2 Icelandic error corpora.

To compare the errors of L2 speakers to native speakers, a contrastive analysis was conducted between the L2 error corpus and the general Icelandic Error Corpus (Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, Þ., 2020).

As Table 1 demonstrates, the number of errors per 1000 words is significantly higher in L2 texts than in the general corpus, and despite the general corpus being much larger with tenfold total word count, the total number of errors in the L2 corpus is still quite statistically significant.

This is not surprising as learner errors are quite frequent, and particularly on lower proficiency levels the text can be so convoluted and inaccurate that making revisions proved to be a challenge as sometimes entire sentences needed to be rewritten for the text to be semantically coherent. However, it must be

noted that the learner error corpus contains significantly fewer and less genre-diverse texts so this may not be reflective of L2 users as a population.

The analysis also sheds light on a significant disparity in frequency of certain error categories and subcategories in L2 Icelandic compared to L1 errors. The most frequent error category in the L2 corpus is grammar, which accounts for almost half of all errors (43.57%). In comparison, the category grammar accounts for only 11.8% in the general Icelandic Error Corpus. The punctuation (12.14%) and wording (11.63%) subcategories are most prominent in second tier, where *wording* is also a specific error code with highest frequency (unsurprisingly, as many otherwise unsorted errors connected to choice of words tend to fall under it). Each other error category comprises 5% or less of total errors. Depicted in Table 2 are all error codes that appear only in the L2 corpus, 30 of which are within the grammar category.

Grammar	
case-verb	case-prep
genitive	pro-inflection
act4mid	missing-sub
tense4perfect	missing-fin-verb
case-collocation	act4pass
extra-sub	mid4act
extra-dem-pro	missing-dem-pro
v3-subordinate	numeral-inflection
perfect4tense	missing-obj
adj4noun	noun4adj
mid4pass	case-adj
pass4mid	pass4act
passive	geta
extra-fin-verb	extra-prep
extra-munu	syntax-other
Orthography	
wrong-symbol	
Vocabulary	
context	interr-pro
though	þar4það

Table 2: Error codes that appear only in the L2 error corpus

These errors mostly involve case government (*case-verb*, *case-collocation*, *case-prep*, *case-adj*) as it is not intuitive in the language learning process which case is governed by a certain preposition or verb, as well as the use of grammatical voice, and inflectional errors in closed word classes. Inflectional errors in nouns or verbs are also among the most common errors but are also prominent in the L1 corpus. Fixed word order in Icelandic is not intuitive for the learner either which created two additional error subclasses within syntax. Another very specific error type for L2 in the lexical category is *context* – an incorrect word chosen for the specific context often prompted by a literal dictionary translation.

The frequency of error codes was ranked to identify to which extent subclasses differ in frequency between the corpora. When the frequencies of multiple error codes were identical, they were ranked equally. If the error code does not appear in a corpus, the rank is by default higher by one than the total number of ranks. The relative rank (Δ rank) between the corpora was calculated for each error code. A high number indicates a large difference in ranks between corpora for an error code, and a low number indicates similar rankings.

Table 3 shows that the two types of error that are ranked among highest in both corpora are *wording* and *nonword* error, the latter being possibly a simple typing error or an attempt to write a word form that

Error Codes	Main Category	Subcategory	Rank L1	Rank L2	Δ rank
wording	style	wording	1	1	0
nonword	orthography	nonword	3	3	0
date-abbreviation	orthography	punctuation	99	99	0
extra-conjunction	vocabulary	insertion	32	33	1
comma4colon	orthography	punctuation	89	90	1

Table 3: Error codes with most similar rankings between the corpora.

does not exist, whereas the former is the most general error type which includes any type of formulating a phrase or a clause in a wrong way, and is often combined with other error types.

4.2 Errors by Proficiency Level and Length of Residence

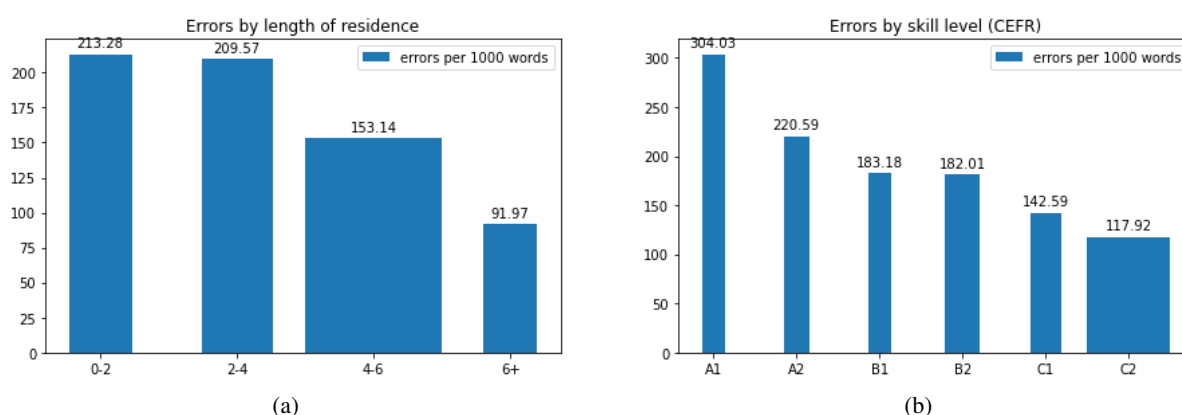


Figure 2: Errors per 1000 words based on length of residence and proficiency level.

Factors that are generally considered in evaluating the interlanguage development are the length of study of the target language as well as the level of interaction and use of the language (does the learner live in the country where the language is spoken, how much are they exposed to the language daily and through which outlets). However, recent research has advised against relying on criteria that assign proficiency by length of study in a language learning program and suggest rather that each learner's production be individually assessed (Thewissen, J., 2013).

Level	Files	Total words	Total errors	Errors/1000w
A1	19	7,759	2,359	304.03
A2	19	11,900	2,625	220.59
B1	12	12,900	2,363	183.18
B2	11	19,504	3,550	182.01
C1	10	22,617	3,225	142.59
C2	14	72,785	8,583	117.92

Table 4: Total number of files, words, errors and errors per 1000 words per skill level.

In this corpus, the proficiency level that learners achieve mostly correlates with the time they have spent residing in Iceland. For 31 of the submitted texts, the author started learning the language before arriving to the country, whereas for the remaining texts the learner started learning the language imme-

diately or some time after starting to live in Iceland and is likely to have not had much contact with the language before commencing the formal learning process. The two graphs in Figure 2 show the number of errors per 1000 words based on length of residence and the proficiency level and show a downward curve which was to be expected as SLA progresses. The width of the bars indicates the total number of words within the given category, notably highest in 4-6 years of residence and level C2, as was briefly mentioned in section 2.1 and laid out in Table 4 (as received texts on this level were mostly longer essays or entire theses). An interesting development is that the downward trend is not as sharp between levels B1 and B2, and in fact the frequency of *wording* and *nominal-inflection* errors increases, which is in line with the typical regression point mentioned in section 2, that the probability of some types of errors increases rather than decrease between levels B1 and B2. Even though this type of regression is expected and observed through SLA research, it needs to be stated that there are no strict guidelines and defined grammatical and lexical requirements that learners need to reach to officially be on a certain level on the CEFR scale for Icelandic. Therefore, it is entirely possible that at the beginning of their second year of the BA in Icelandic as a second language, not all students have reached level B2 and there is likely some overlap in the labelling of texts between these two CEFR levels.

Finally, we should also keep in mind that the general error corpus has the average of 44 errors per 1000 words whereas the average for the highest level (C2) and longest dwelling (more than 6 years) is 117 and 91 respectively which shows that non-native speakers are more than twice as likely to make mistakes even as they approach near-native competence as much as possible. However, the nature of errors changes over time.

Table 5 shows the frequency of the (5) most common errors per proficiency level. *Nominal-inflection* is an error type that consistently ranks among the most frequent. This is not surprising, and as Þorvaldsdóttir and Garðarsdóttir point out in their research (Garðarsdóttir, M., and Þorvaldsdóttir, S., 2020), acquisition of cases is a slow process with many typical points of overgeneralization (subject as nominative, object as accusative, atypical subject as dative etc.), and the results indicate that the learners assume the so-called structural case later than thematic case with the most atypical idiosyncratic case being the latest and most reluctantly accepted which in most cases is the genitive case. A common source of case errors also comes from the previously mentioned case governance which also ranks high on all levels. For example, L2 users will commonly misuse a phrasal verb and instead of interpreting it as a verb clause they would take the preposition as part of a prepositional clause with the following noun and apply the case that preposition governs (so [*leysa af*] + accusative becomes *leysa* + [*af* + dative]).

Although the general wrong-choice-of-words error, *wording*, ranks consistently highest (and as was shown, is also the most frequent error type in the general corpus), the occurrence of the more particular choice of word error, *context*, drops between levels, being in 4th place on A2 and dropping to the 7th place on level C1. Another common error type, *nonword*, is likely to not be a competence error but accidental or caused by over- or underuse of specific Icelandic accented vowels. However, its frequency being significantly higher on the lowest proficiency level does to some extent stem from overgeneralization, e.g. assigning a wrong gender to a noun creating an incorrect inflection form, or conjugating an irregular verb as regular. Other noteworthy error types are *ind4sub* / *sub4ind* which track the incorrect use of the subjunctive mood and rise in frequency as proficiency level increases (particularly *ind4sub*, as learners tend to overuse the indicative mood). This is not surprising, as beginner and lower intermediate learners use simpler sentence structures and do not learn about grammatical mood until later on in the learning process, and this type of grammatical error is among the most common for native speakers as well.

Lastly, there is a number of ambiguous cases where it is not clear whether an error is a spelling or a grammatical error and in these cases it is hard to estimate the author's intention. Here the error would be categorized based on the overall analysis of the given text and the skill level and its linguistic expectations, i.e. if the text has repeated unambiguously inflectional errors, the error in question would most likely be categorized as such, whereas if the grammatical correctness of the text overall is high, it would be interpreted as a spelling-orthography related error.

5 Conclusion

This paper introduces the Icelandic L2 Error Corpus, the first learner error corpus for Icelandic, which is a collection of texts written by users of Icelandic as a second language. The majority of the texts are student essays submitted by students in the Icelandic as a second language program at the University of Iceland. The texts have been manually annotated for errors based on an error tagset previously built for the general Icelandic Error Corpus based on native speaker texts. First results of two CIA approaches are also presented, first comparing the L2 corpus with the general corpus and second analysing the L2 error corpus focusing on proficiency level.

Error Codes	Category	Subcategory	Freq	Errors/1000w
A1				
wording	style	wording	236	30.42
nominal-inflection	grammar	inflection	115	14.82
nonword	orthography	nonword	91	11.73
missing-period	orthography	punctuation	76	9.79
extra-word	vocabulary	insertion	72	9.28
A2				
wording	style	wording	260	21.85
nominal-inflection	grammar	inflection	185	15.55
wrong-prep	grammar	prep	97	8.15
context	vocabulary	lexical	91	7.65
extra-comma	orthography	punctuation	88	7.39
B1				
wording	style	wording	254	19.69
nominal-inflection	grammar	inflection	93	7.2
extra-word	vocabulary	insertion	83	6.43
extra-comma	orthography	punctuation	83	6.43
def4ind	grammar	definitiveness	81	6.27
B2				
wording	style	wording	558	28.61
nominal-inflection	grammar	inflection	202	10.36
extra-word	vocabulary	insertion	87	4.46
nonword	orthography	nonword	84	4.31
missing-word	vocabulary	omission	84	4.31
C1				
wording	style	wording	453	20
nonword	orthography	nonword	130	5.75
ind4def	grammar	definitiveness	108	4.77
nominal-inflection	grammar	inflection	105	4.64
agreement-concord	grammar	agreement	99	4.38
C2				
wording	style	wording	802	11.02
nominal-inflection	grammar	inflection	508	6.98
nonword	orthography	nonword	309	4.24
wrong-prep	grammar	prep	296	4.07
extra-comma	orthography	punctuation	249	3.42

Table 5: Most common error codes by proficiency level in the L2 corpus

At this point, the corpus consists of 15,571 revision spans and 22,705 categorized error instances. When compared to the L1 error corpus for Icelandic which has more than a million words and 44,261 revision spans, the overall number of errors is only slightly less than half of the number of errors in the general corpus, which is both valuable for research and analysis and developing the correction tool, and indicates a great distinction in the frequency of errors L2 and L1 users make in written form, as was further shown in our analysis. It should be noted that a total of 85 texts by 36 learners with only 15 first languages, and at the moment of writing, the immigrant population in Iceland uses more than 100 different native languages, might not be fully representative of the L2 community in Iceland (and we hope the further expansion of the corpus will provide more diversity).

The preliminary results show a large disparity in the quantitative distribution of errors in the Icelandic L2 Error Corpus and the general Icelandic Error Corpus. This disparity relates to both the occurrence of different error categories, where grammar related errors are 4 times more prominent in the L2 corpus, and the total error rate, which is 3 times as high for the L2 corpus compared to the native speaker referent. Moreover, it is still more than twice as high when the L2 speakers have reached the highest proficiency level and dwelled in the country for more than 6 years. The L2 vs. L2 analysis also yielded interesting yet predictable results, showing the downwards trend of error occurrences as the learner's proficiency and their length of residence in Iceland increases, with a notable slight increase in certain types of errors between levels B1 and B2.

Learner error corpora are important for shedding light on learner interlanguage which can aid the development of various automatic language correction tools and teaching materials and also provide insight into how and in which order certain grammatical and lexical categories are acquired and internalized. Thus we hope to further expand this corpus to provide more possibilities to analyse various features of learner language that could not be covered so far due to the limited size of the sample and its lack of diversity of highlighted linguistic features. With the expansion of the corpus, it has potential to become an important asset for learning Icelandic.

References

- Cummins, J. 1991. *Interdependence of first- and second-language proficiency in bilingual children*, page 70–89. Cambridge University Press.
- Díaz-Negrillo, A., and Fernández-Domínguez, J. 2006. Error tagging systems for learner corpora. *Revista española de lingüística aplicada, ISSN 0213-2028, Vol. 19, 2006, pags. 83-102*, 19, 01.
- Friðriksdóttir, S. R., and Ingason, A. K. 2020. Disambiguating confusion sets in a language with rich morphology. In *Proceedings of ICAART 12 (International Conference on Agents and Artificial Intelligence)*.
- Gamon, M., Leacock, C., Brockett, C., Dolan, W., Gao, J., Belenko, D., and Klementiev, A. 2013. Using statistical techniques and web search to correct ESL errors. *CALICO Journal*, 26:491–511, 01.
- Garðarsdóttir, M., and Þorvaldsdóttir, S. 2020. A processability approach to the development of case in L2 Icelandic. *Language, Interaction and Acquisition A cross-theoretical and cross-linguistic perspective on the L2 acquisition of case systems / Lacquisition de systèmes casuels en L2 : des études à travers plusieurs théories et langues*, 11(1):68–98.
- Golding, A., and Roth, D. 1999. A winnow-based approach to context-sensitive spelling correction. *Machine learning*, 34(1-3):107–130.
- Granger, S., 2008. *Learner Corpora in Foreign Language Education*, pages 1427–1441. 01.
- Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. 2009. Context-sensitive spelling correction and rich morphology. In Jokinen, K. and Bick, E., editor, *Proceedings of NODALIDA 2009*, pages 231–234.
- Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I. 2021. The Icelandic L2 error corpus (IceL2EC) version 1.1. CLARIN-IS.
- Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, Þ. 2020. The Icelandic error corpus (IceEC). version 1.0.

- Piccardo, E., Goodier, T., and North, B. 2018. *Council of Europe (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe Publishing., 01.
- Pienemann, M., Di Biase, B., Kawaguchi, S., and Håkansson, G. 2005. Processing constraints on L1 transfer. *Handbook of bilingualism: Psycholinguistic approaches*, pages 128–153, 01.
- Pienemann, M. 2011. *Studying processability theory : an introductory textbook / edited by Manfred Pienemann, Jorg-U. Kessler*. Processability approaches to language acquisition research teaching (PALART), v. 1. John Benjamins Pub. Co., Amsterdam ;.
- Sigmundsson, S. 1987. Íslenska í samanburði við önnur mál. *Íslenskt mál og almenn málfræði*, 9.
- Thewissen, J. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1):77–101.
- Ólafsson, G. H. 2016. Grammar and linguistic structures at level A1 of Icelandic. Master's thesis, University of Iceland, Unpublished, 6.
- Þorvaldsdóttir, S., and Garðarsdóttir, M. 2013. Fallatleinkun í íslensku sem öðru máli. *Milli mála: Tímarit um erlend tungumál og menningu*, 5:45–73.

The TEI-based ISO Standard ‘Transcription of spoken language’ as an Exchange Format within CLARIN and beyond

Hanna Hedeland

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

Thomas Schmidt

Research and Infrastructure Support
Universität Basel, Switzerland
th.schmidt@unibas.ch

Abstract

This paper describes the TEI-based ISO standard 24624:2016 ‘Transcription of spoken language’ and other formats used within CLARIN for spoken language resources. It assesses the current state of support for the standard and the interoperability between these formats and with relevant tools and services. The main idea behind the paper is that a digital infrastructure providing language resources and services to researchers should also allow the combined use of resources and/or services from different contexts. This requires syntactic and semantic interoperability. We propose a solution based on the ISO/TEI format and describe the necessary steps for this format to work as an exchange format with basic semantic interoperability for spoken language resources across the CLARIN infrastructure and beyond.

1 Introduction

Today, the CLARIN infrastructure is well established across Europe, comprising a network of centres providing a vast number of digital resources and services. Since an increasing number of funders require researchers in the humanities and social sciences to deposit their data for reuse, the collections of digital resources hosted within CLARIN are growing steadily. Following the digital turn, the use of CLARIN’s tools and services for manual and automatic analysis has also become a relevant option for research projects from various disciplines. An ideal scenario would allow researchers to use and freely combine data and tools or services from different CLARIN centres and contexts across the infrastructure. This, however, is still possible only for smaller sets of resources – large scale interoperability remains a desideratum. Unlike early digital corpora created by pioneering corpus linguists, digital language resources today seldom fit into the traditional view of language data as ‘natural running text’ or ‘a single stream of tokens’. This is particularly true for spoken or multi-modal resources, which are at the same time no longer a rare exception in the resource landscape.

In this paper, a TEI-based ISO standard for the representation of spoken language transcription will be introduced and its current and future relevance for CLARIN and related contexts will be discussed. After this introduction we will provide an overview of tools and services which are currently available to work with that standard in creating, enriching and publishing spoken language data.

2 A Standard for Spoken Language Transcription?

2.1 Interoperability of Existing De-Facto Standards and Tool Formats

One reason for the heterogeneity of spoken language corpora is the existence of several widely used tool formats. ELAN (Sloetjes, 2014), Praat (Boersma, 2001), CLAN (MacWhinney, 2000), Transcriber (Barras et al., 2001), FOLKER (Schmidt, 2016) and EXMARaLDA (Schmidt and Wörner, 2014) all come with their individual formats, which are, apart from Praat’s TextGrid format, XML-based. These formats are mainly based on similar tier-/time-based data models, i.e. they model transcription as a set of

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

tiers with different characteristics containing different information, and are already to a sufficient extent interoperable – from the syntactic perspective (Schmidt et al., 2009). A file in one format can usually be converted into a file with a representation of the data using another format. There are undoubtedly some limitations regarding conversion scenarios, depending on the varying complexity of data models, where e.g. certain tier hierarchies or associations between annotation elements in ELAN’s EAF format cannot be modelled by the more restrictive data model for Basic Transcriptions (EXB) in the EXMARaLDA system. In these rather rare cases, customised workarounds are still possible.

From a semantic perspective however, interoperability is not that straightforward, since both the set of tiers used and their content vary to a great extent. One solution to this dilemma would be to standardise tiers and tier content. As an example, the CHAT format of the CLAN software, depicted in Figure 1, exactly defines the set of transcription and annotation conventions to be used for common spoken language phenomena, which makes the data easy to process and understand. But researchers are at the same time required to subscribe to theoretical concepts implemented by these conventions, and this is not a good basis for a standard to be used across discipline boundaries.

```

@Begin
@Languages:      eng
@Participants:  CHI Ross Child, FAT Brian Father
@ID:            eng|macwhinney|CHI|2;10.10||||Target_Child|||
@ID:            eng|macwhinney|FAT|35;2.||||Target_Child|||
*ROS:          why isn't Mommy coming?
%com:          Mother usually picks Ross up around 4 PM.
*FAT:          don't worry.
*FAT:          she'll be here soon.
*CHI:          good.
@End

```

Figure 1: The CHAT transcription system defines the units of the transcription, the annotation tiers and the transcript layout.

On the other side of the spectrum, the EAF format of the ELAN software hardly imposes any restrictions on the individual researcher, who is free to define the structure and content of the data format according to her needs. While this promises a perfect fit for the individual research context, data modelling is not trivial and not all variation is semantically relevant. This means that transcripts containing e.g. a basic orthographic transcription, interlinear glosses and a translation into English can be modelled in various ways using different tier types and names, making automatic processing of similar resources difficult since the semantics of the tiers are only documented for humans. It should be noted that ELAN has been providing means to define the semantics of tiers and annotations using external controlled vocabularies or references to ISOcat for many years. The comprehensive evaluation of annotation practices in language documentation corpora presented by von Prince and Nordhoff (2020) shows that this has however hardly been adopted by researchers using the software. This might be related to the proliferation of data categories in ISOcat or simply a matter of lacking awareness of the problem.

2.2 The ISO/TEI Approach to Standardisation and Interoperability

The ISO standard for Transcription of spoken language (ISO/TC 37/SC 4, 2016; Schmidt, 2011) is based on the TEI Guidelines (TEI Consortium, 2021), mainly on the chapter ‘8 Transcriptions of Speech’¹. The idea behind the standard is to find a solution that differentiates between general information that is shared across different research methods and disciplines on the one hand, and information that is theory-dependent (cf. (Ochs, 1979)) and therefore cannot be standardised, on the other. Standardisation can be applied to aspects of the shared reality of spoken conversation, which includes e.g. the modelling of participants and the temporal alignment of their contributions. These aspects, referred to here as macro-structure, are not defined by transcription conventions or other theoretical constructs.

¹<https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>

Unlike many of the widely used transcription tool formats, the ISO/TEI format depicted in Figure 2 is not a pure tier-/time-based format. Instead, it models speaker contributions as a common list of <u> elements. Its structure is thus more similar to written documents. Speaker contributions are often considered to comprise several linguistic units, accordingly <u> elements may contain one or more <seg> elements corresponding to the linguistic units defined by the relevant transcriptions system via @type and @subtype attributes. References to defined speakers and time points are modelled by the attributes @who, @start and @end, with the option to use <anchor> elements for additional alignment in any position². Annotations are by default modelled in a standoff manner by elements in <spanGrp> elements, the annotation level defined by a @type attribute. Annotations can be used with <anchor> elements as in Figure 2 or refer to words, <w> elements, if the text has been tokenised and marked-up accordingly. An additional element <annotationBlock> is used to group the speaker contribution <u> with all annotations referring to it.

```
<text xml:lang="en">
  <timeline unit="s">
    <when xml:id="T0" />
    <when xml:id="T3" />
    <when xml:id="T4" />
    <when xml:id="T1" />
    <when xml:id="T2" />
  </timeline>
  <body>
    <annotationBlock xml:id="AB1" who="#TIM" start="#T0" end="#T1">
      <u xml:id="U1">Excuse me, is <anchor synch="#T3" />this <anchor synch="#T4" />the way out? </u>
      <spanGrp type="pho">
        <span from="#T3" to="#T4">[zis]</span>
      </spanGrp>
    </annotationBlock>
    <annotationBlock xml:id="AB2" who="#TOM" start="#T1" end="#T2">
      <u xml:id="U2">Yes, straight ahead. </u>
    </annotationBlock>
  </body>
</text>
```

Figure 2: A simple example of the transcription macro-structure of the ISO/TEI format.

Below the macro-structure, within the speaker contribution, there are many differences in the precise form of representation for verbal and accompanying non-verbal elements and features across transcription systems. We will refer to this level, which may also contain widely recognised linguistic units such as words, as the micro-structure. The differences between the representations used in various transcription systems are partly due to important reflections of theoretical differences, but in other cases the syntactic differences resulting from the choices of transcription symbols do not reflect any semantic differences, and in some cases syntactic or symbolic identity obscures semantic differences. Figure 3 shows the traditional printed representation of the same speaker contribution using two different transcription systems.

MJ: I ((cough)) see a door. I (0.3) want to paint it (black/blue).

MJ[v] I ((cough)) see a door. I ((0,3s)) want to paint it (black).
 MJ[k] (blue)

Figure 3: The same speaker contribution transcribed according to two different transcription systems; GAT (Selting et al., 1998) (above) and HIAT (Rehbein et al., 2004) (below)

To the human reader, the similarities are striking and the slight differences in the representation of identical phenomena are easily deciphered. Both transcription systems use double parentheses to represent non-verbal and non-phonological elements, the green highlighting of the ‘((cough))’ was therefore added to this example to indicate syntactic and semantic identity. The short (0.3 seconds) pause and the uncertainty regarding which colour (black or blue) will be used to paint the door share the same semantics

²Owing to performance reasons and ease of processing, the ZuMult project (cf. Section 4.3) uses ID/IDREFs instead of XPointers for pointing between elements.

but are syntactically different, the added highlighting is yellow. The uncertain part is even structurally different, since the HIAT system (below in Figure 3) requires the alternative interpretation to be transcribed in an additional tier for comments ('k') below the main transcription tier. The two full stops highlighted in red are on the contrary syntactically identical, but their semantics differ, since the two transcription systems use this symbol to denote different types of units within the speaker contribution.

It is possible to represent this example in the ISO/TEI format without taking the transcription conventions into account. In Figure 4 this has been done for the same example with the GAT version above and the HIAT version below. The same similarities and differences still apply and the structural difference in the representation of uncertainty is encoded once through symbols in the text of the speaker contribution for GAT (above in Figure 4) and once as an annotation of the uncertain part for HIAT (below in Figure 4). With this representation of the data in the same format, syntactic interoperability has been achieved. Reliable automatic processing or querying of the content of this type of data across collections using different transcription systems still remains difficult, since there is no semantic interoperability on this level.

```

<u who="#MJ" start="#T0" end="#T2">
  <seg type="contribution">
    I ((cough)) see a door.
    <anchor synch="#T1"/>
    I (0.3) want to paint it (black/blue).
  </seg>
</u>

<u who="#MJ" start="#T0" end="#T4">
  <seg type="contribution">
    I ((cough)) see a door.
    <anchor synch="#T1"/>
    I ((0,3s)) want to paint it <anchor synch="#T2"/>(black)<anchor synch="#T3"/>.
  </seg>
</u>
<spanGrp type="k" subtype="time-based">
  <span from="#T2" to="#T3">(blue)</span>
</spanGrp>

```

Figure 4: The examples can be represented in the ISO/TEI format without using the implicit information of the transcription conventions.

```

<u who="#MJ" start="#T0" end="#T2">
  <seg type="intonation-phrase" subtype="falling">
    <w>I</w><vocal><desc>cough</desc></vocal><w>see</w><w>a</w><w>door</w>
  </seg>
  <anchor synch="#T1"/>
  <seg type="intonation-phrase" subtype="falling">
    <w>I</w><pause dur="PT0.3S"/><w>want</w><w>to</w><w>paint</w><w>it</w>
    <unclear><choice><seg><w>black</w></seg><seg><w>blue</w></seg></choice></unclear>
  </seg>
</u>

<u who="#MJ" start="#T0" end="#T2">
  <seg type="utterance" subtype="declarative">
    <w>I</w><vocal><desc>cough</desc></vocal><w>see</w><w>a</w><w>door</w>
  </seg>
  <anchor synch="#T1"/>
  <seg type="utterance" subtype="declarative">
    <w>I</w><pause dur="PT0.3S"/><w>want</w><w>to</w><w>paint</w><w>it</w>
    <unclear><choice><seg><w>black</w></seg><seg><w>blue</w></seg></choice></unclear>
  </seg>
</u>

```

Figure 5: When encoded using the ISO/TEI format, the partly identical meaning of the different transcription symbols becomes explicit and only the theory-dependent differences of Figure 4 remain.

Semantic interoperability can be achieved through standardisation, though while some aspects of the micro-structure can be standardised, such as the existence of pauses and (possibly) non-verbal behaviour, the detailed choices regarding e.g. a set of relevant different pause durations or the descriptions of non-verbal behaviour have to correspond to the theory-dependent transcription system currently in use. The same is true for the details of the segmentation into linguistic units in <seg>s, which usually differs according to the linguistic level used as the basis. Allowing for controlled variation within this area makes it possible to represent data created with different transcription systems using the same standard format. In Figure 5 the micro-structure has been parsed according to the different transcription systems during the conversion process and a common representation of shared phenomena – the word and non-word tokens, the pause, and the uncertainty with the alternative interpretation – has been achieved. It has also become possible to explicitly express the different semantics of the units below the speaker contribution, i.e. the different meaning of the full stop in the two transcription systems, through the use of <seg> elements with @type and @subtype attributes, in this case intonation phrases based on interactional prosody for the GAT system and utterances based on the pragmatics level for the HIAT system. This type of conversion results in transcription data that is semantically interoperable where this is possible and for which semantic and theory dependent differences become explicit and machine-readable.

3 Acceptance of ISO/TEI and Related Formats in CLARIN

Within CLARIN, centres are not bound to accept or support particular formats. In accordance with the requirements of the CoreTrustSeal (CoreTrustSeal Standards and Certification Board, 2019), which is a prerequisite for the certification of CLARIN B centres (cf. (Wittenburg et al., 2019)), all centres do however provide information about accepted file formats for resource deposits. Some centres have compiled individual lists for this purpose and others still refer to one of several older general lists and overviews of standards and recommendations for CLARIN³. While these lists pre-date the ISO/TEI format, they all include TEI as a general recommendation. At the time of writing, seven B centres point to such external information⁴.

The CLARIN Standards Committee has been gathering information on the recommendations on standards and formats actively issued by individual (mainly B) centres and made this information available on their web page⁵ and as the basis for the relaunch of the CLARIN Standards Information System (SIS)⁶. A brief assessment of this information can provide insights into the current and potential support for the ISO/TEI standard within CLARIN. For this paper, the Standards Information System and the original recommendations given by individual centres were surveyed. Since the transformation from the various centres' individual recommendations into the SIS might be a source of inaccuracy, the original documents and websites were revisited for centres that have not validated and confirmed their SIS information. As not all centres accepting data deposits provide detailed individual recommendations yet, the picture is however still not complete. Since there is also no consistent and reliable information on the general types of resources a centre accepts nor on specific restrictions e.g. regarding languages or time periods, negative results cannot really be interpreted in the sense of lacking acceptance for ISO/TEI or related formats, since the centre might not accept resource types for which ISO/TEI is a relevant format.

Nevertheless, of the centres that provide their own preferences and recommendations, three groups with respect to ISO/TEI support can be distinguished. According to validated information of the SIS and the centres' original recommendations at the time of writing, four B centres already recommend ISO/TEI explicitly⁷. These are the CLARIN.SI Language Technology Centre, The Language Bank of Finland (FIN-CLARIN), the Hamburg Centre for Language Corpora (HZSK) and the Leibniz-Institut für Deutsche Sprache (IDS). In addition to the information from certified B centres, the centres TOols for LANGuage (ORTOLANG) and Language Archive Cologne (LAC), which are both participating in

³Such resources are e.g. <https://www.clarin.eu/faq/what-standards-are-recommended-clarin> or <https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf>

⁴cf. <https://github.com/clarin-eric/standards/issues/14>

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


⁶<https://standards.clarin.eu/sis/>

⁷cf. <https://standards.clarin.eu/sis/views/view-format.xq?id=fTEISpoken>

ISO/TEI Transcriptions of Spoken Language

Abbreviation: TEISpoken

Identifiers:

Type	Id
SIS ID	fTEISpoken 

Media type(s):

- application/tei+xml;format-variant=tei-iso-spoken
- application/tei+xml;format-variant=tei-iso-spoken;tokenized=[0,1]

File extension(s): .tei

Format family: TEI

Recommendation:

Clarín Centre	Domain	Level	Comments
HZSK	Audiovisual Annotation	recommended	
ZIM	Audiovisual Annotation	recommended	
IDS	Audiovisual Annotation	recommended	
FIN-CLARIN	Audiovisual Annotation	recommended	
CLARIN.SI	Audiovisual Annotation	recommended	

Description:

This format is a serialization of the [ISO/TEI Transcriptions of Spoken Language](#).

ISO/TEI transcriptions of spoken language will be identified by the MIME type `application/tei+xml;format-variant=tei-iso-spoken`. A parameter `tokenized=0/1` can be added to indicate whether (=1) or not (=0) the respective TEI file is tokenized (i.e. has `<w>` markup).

For more information, see [Thomas Schmidt, "A TEI-based Approach to Standardising Spoken Language Transcription", Journal of the Text Encoding Initiative \[Online\], Issue 1 | June 2011, Online since 08 June 2011, connection on 21 September 2021. URL: <http://journals.openedition.org/jtei/142>; DOI: <https://doi.org/10.4000/jtei.142>](#)

Please feel welcome to supply the description of this format file via GitHub: either as an [issue report](#), or as a [pull request](#) after forking or browsing the [code](#) under the 'formats' branch.

[\[suggest a fix or extension\]](#)

Keywords: annotation format, corpus encoding

Figure 6: Information on the ISO/TEI format in the CLARIN Standards Information System (SIS).

CLARIN knowledge centres and aiming for B Centre status, also explicitly recommend the ISO/TEI format for incoming deposits.

Forming a second group, further centres recommend TEI, and thus implicitly ISO/TEI, though this variant is not explicitly mentioned⁸. Among these are the Austrian Centre for Digital Humanities and Cultural Heritage - A Resource Centre for the HumanitiEs (ACDH-ARCHE), Eberhard Karls Universität Tübingen (EKUT), the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), The CLARIN Centre at the University of Copenhagen (CLARIN-DK-UCPH), the ZIM Centre for Information Modelling (ZIM) and the Meertens Instituut/HuC (MI) (which only includes XML in the list, but refers to TEI as an example). The centre Collections de corpus oraux numériques (COCOON) also recommends TEI, is however not a certified B centre. As noted above, all centres referring to existing CLARIN documents also in effect recommend TEI without further restrictions.

The third group is the most interesting, since these centres explicitly recommend other widely used formats and not ISO/TEI. The CMU-TalkBank (CMU) recommends CHAT (only), MPI for Psycholinguistics (MPI-PL) recommends CHAT too, though in addition to EAF and Praat, which are in turn also

⁸cf. <https://standards.clarin.eu/sis/views/view-format.xq?id=fTEI>

recommended by the Bayerisches Archiv für Sprachsignale (BAS). Both Praat and EAF can be converted into the ISO/TEI format with dedicated software as described in (Schmidt et al., 2017), and this also applies to CHAT data that passes the data quality and consistency tests in CLAN. Still, the ISO/TEI format seems to be of little relevance to these four centres, presumably because of strong traditions and eco-systems around specific formats for specific types of resources and research areas. Furthermore, the LINDAT/CLARIAH-CZ centre, which does not give explicit recommendations on formats to depositors, now hosts the TEI-based TEITOK system (Janssen, 2021), which includes both a search engine, visualisation and editing functionality and has many features for spoken language. Since this TEI variant is interoperable with e.g. EXMARaLDA and EAF through a set of scripts, interoperability between the TEITOK and ISO/TEI formats is also feasible.

As expected, TEI, the ISO/TEI format, and formats that can be converted into the ISO/TEI formats are often recommended for resource deposition across the infrastructure. A more systematic approach towards the description and dissemination of format recommendations would facilitate further steps towards enhanced interoperability for transcription data in CLARIN. The Standards Information System can now be used to manage and analyse the relevant information as provided by the centres.

4 Tools and Services for ISO/TEI within and beyond CLARIN

Whether or not a new standard is widely adopted crucially depends on how well it interoperates with existing tools and methods. Ideally, researchers can continue working with established workflows and will profit from additional benefits because these workflows are becoming standard-compliant. The ISO/TEI standard was defined with this practical goal in mind. In what follows we will look at different stages of the research data lifecycle for spoken language corpora, explaining and illustrating how existing tools and methods interoperate with the standard.

4.1 Data Creation (Transcription)

Among the existing, widely used tools for transcription (see above), the EXMARaLDA Partitur-Editor and FOLKER/OrthoNormal provide the most direct interoperability with ISO/TEI. The tools continue to write their tool specific format, but now have an additional option for exporting ISO/TEI. In the case of the Partitur-Editor, the export can be configured to use different algorithms for segmenting transcribed text into word and non-word tokens (such as pauses or descriptions of non-verbal behaviour) according to different transcription systems (see Figure 7). The Partitur-Editor can also import files in the ISO/TEI format. Since the internal tool format does not represent tokens and other parts of the micro-structure, this is strictly speaking a lossy transformation. The information, however, can be automatically reconstructed from implicit information during the corresponding export process.

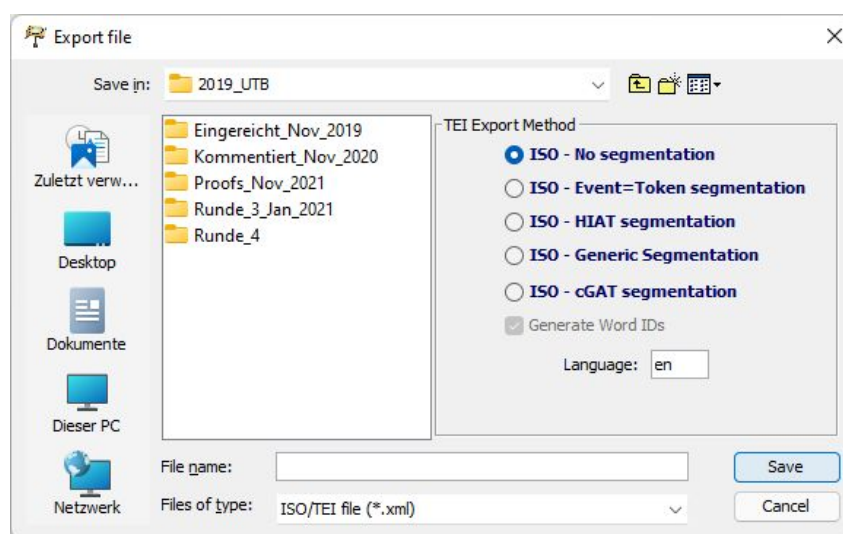


Figure 7: ISO/TEI export dialog of the EXMARaLDA Partitur-Editor.

The other transcription tools mentioned above (i.e. ELAN, Transcriber, CLAN and Praat) do not (as yet) provide direct means of importing or exporting ISO/TEI. The conversion can, however, be achieved via the EXMARaLDA Partitur-Editor (which has import filters for all of the formats), via TEI-Drop, a dedicated tool for that purpose, or via web-services (Schmidt et al., 2017).

4.2 Data Enrichment (Annotation)

Since the creation of the ISO/TEI standard, the format has been used as the basis for enhanced interoperability with existing annotation tools and services. In many cases, this was software created on the basis of data models or notions of written language. Since the ISO/TEI standard is a TEI-based format, it shares a common core with TEI variants used for written language data and thus facilitates interoperability across the spoken and written modality. For instance, the development of WebAnno-MM (Remus et al., 2019) as an extension for audiovisual and transcription data in the ISO/TEI format allows manual annotation with a wider textual focus than transcription tools offer, and also more complex types of annotations such as tree or chain annotations. The original user interface for annotation tasks and the score visualisation for transcription data are shown in Figure 8.

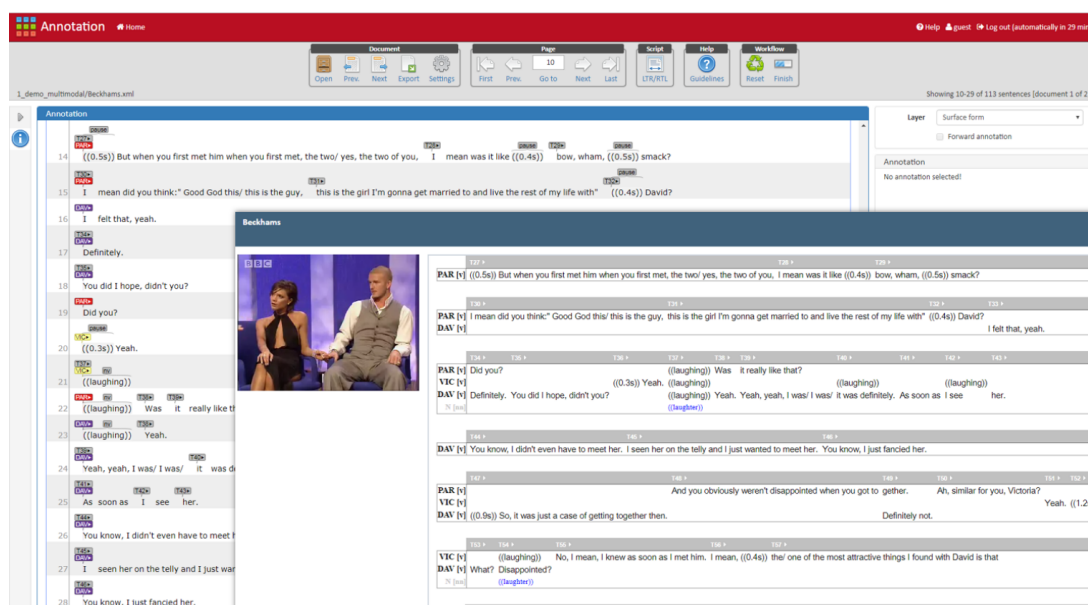


Figure 8: Annotation and multimedia transcript score view in WebAnno-MM.

For automatic annotation, the converters described above were integrated into the WebLicht SOA (Hinrichs et al., 2010) of CLARIN-D, thus enabling the use of various services from all German centres. Initially, this meant another mapping to formats and services for written data (internally, TCF, see

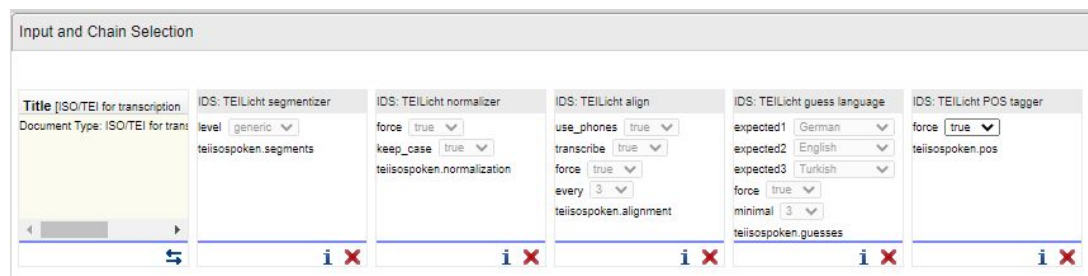


Figure 9: An ISO/TEI annotation chain defined in WebLicht.

(Schmidt et al., 2017)), but services adapted to spoken language data based directly on the ISO/TEI format have now also been developed (Fisseni and Schmidt, 2020) and can improve results where the

linguistic characteristics of spoken and written language differ to a great extent. A sample processing chain is shown in Figure 9. The speech data web services provided by the BAS (Kisler et al., 2017) have been able to import and export ISO/TEI data since version 2.36 of January 2020.

4.3 Data Publication and Analysis (Dissemination)

Based on the ISO/TEI format, the project ZuMult has developed new web-based functionality for both visualisation and browsing of spoken language corpora within qualitative approaches and for complex querying and analysis⁹. Query is based on an extension of the MTAS system (Brouwer et al., 2017) which can generate Lucene indices directly from the ISO/TEI XML files. Users can thus be provided with very powerful and efficient querying possibilities in CQP (Frick and Schmidt, 2020). Visualisation uses various XSL transformations to generate, directly from the ISO/TEI XML file, configurable displays of the transcript (in HTML), a density viewer (in SVG) and configurable video subtitles (in VTT) all of which are synchronised with each other and with the underlying audio or video (see Figure 10).

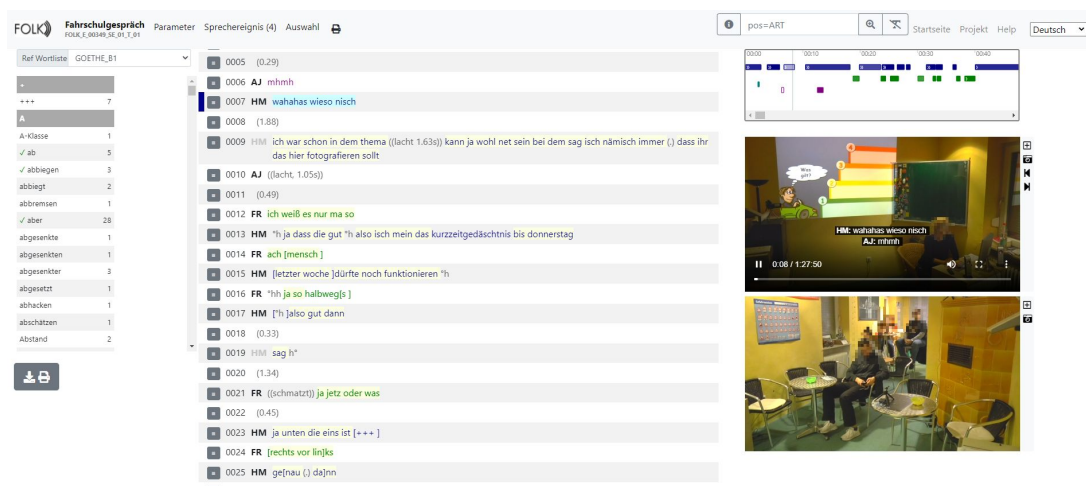


Figure 10: Different visualisations of an ISO/TEI transcript, integrated and synchronised in the ZuViel tool of the ZuMult project.

Another corpus analysis platform that now supports the ISO/TEI format is Tsakorpus (Arkhangelskiy et al., 2019), which is one use case for ISO/TEI within the long-term project INEL in Hamburg (Ferber and Jettka, 2020). A project in the related field of language documentation, the international (French/German) DoReCo project (Paschen et al., 2020), developed the Multitool¹⁰ that can generate ISO/TEI as a distribution format for resources in various languages and tool formats. The ISO/TEI standard is also used as a pivot format for different tool formats in the tool TEICORPO¹¹ developed at the CLARIN K centre CORLI to facilitate data sharing and long-term preservation (Parsse et al., 2020). Since the main aim is a direct lossless conversion from the ELAN, Praat, Transcriber and CHAT formats, the work is complementary to the existing solutions based on the EXMARALDA system. The conversion solutions developed at CORLI also focus on the macro-structure and TEI-conform means of representing arbitrary tier structures found in tool formats of varying complexity without attempts to map micro-structure information systematically.

5 Discussion

The development of interfaces between the ISO/TEI standard and various existing tools and services has shown that this is not only feasible, but also efficient using the ISO/TEI standard as a pivot format. This is important since software development and maintenance is usually the bottleneck in the development

⁹<http://zumult.ids-mannheim.de/ProtoZumult/index.jsp>

¹⁰<https://github.com/DoReCo/multitool>

¹¹<https://ct3.ortolang.fr/teicorpo/>

of the infrastructure. As (Parisse et al., 2020) point out, researchers also need to continue using tools they are familiar with. The ISO/TEI format could enhance interoperability for spoken language resources in CLARIN, especially since the already mentioned centres CORLI, LINDAT and IDS, and parts of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD), already actively use TEI for spoken data. Using a TEI variant to achieve interoperability has also proven successful in the case of parliament corpora (Erjavec et al., 2022). By using a TEI-based format for spoken data, apart from the proximity to more familiar written language data models on the textual level, interoperability on the metadata level could also be facilitated. With the TEI header, there is also a common structure for a core set of relevant contextual information on the setting and the participants, e.g. for analyses within virtual collections. Since TEI is used and extended in many contexts, there are also existing conventions for basic token-based linguistic annotation (Bański et al., 2018) and a common approach for the integration of the W3C standard RDFa is being developed (Chiarcos and Ionov, 2019) to tackle the issue of strict linked data requirements, which are also relevant for the interoperability aspects of the FAIR principles (Wilkinson et al., 2016).

Though conversion is already possible for widely used tool formats, as pointed out above, only features of the macro-structure are strictly defined by the ISO/TEI standard, and only syntactic interoperability is to some extent simple to achieve. For semantic interoperability, the tier structure, the annotation levels and schemas and the conventions for transcription – the micro-structure – also need to be made explicit and machine processable to allow for tokenisation and structural mark-up. This means that a conversion into the ISO/TEI format is not only a question of interoperability with a standard, but at the same time a process of FAIRification, of defining the semantic model of the data, making it more transparent and increasing the number and types of possible re-use scenarios. Creating digital language resources that are FAIR according to the well-known principles is a great, and often somewhat abstract, challenge for CLARIN and its users. We suggest that the adoption of the ISO/TEI standard with its basic semantics and the corresponding conversion scenarios as a way of assessing digital language resources could not only improve interoperability across resources, but also increase their general FAIRness. By using TEI as a common format and settling for answers to the question of machine-readable annotation documentation (Chiarcos et al., 2020) CLARIN could help foster a culture of data documentation required for interoperable and truly FAIR infrastructures for both humans and machines.

6 Conclusion

As this paper has tried to demonstrate, TEI-based standardisation for a sufficiently well-specified domain can make a contribution towards improved syntactic and semantic interoperability in a landscape where different tool-specific formats are already established. Although many issues still remain to be solved, we think that this approach is the most concrete and pragmatic that can be realised in a heterogeneous context such as CLARIN. The ISO/TEI standard, in this sense, is both a technical basis for data exchange in the ‘real world’ and a conceptual model for thinking about farther-reaching standardisation. Adopting such standard proposals as preferred formats of CLARIN centres can further help to consolidate such common ground.

References

- Arkhangelskiy, T., Ferger, A., and Hedeland, H. 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.
- Bański, P., Haaf, S., and Mueller, M. 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).
- Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22. Speech Annotation and Corpus Tools.

- Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glott International*, 5(9/10):341–345.
- Brouwer, M., Brugman, H., Kemps-Snijders, M. 2017. MTAS: a solr/lucene based multi tier an-notation search solution. In *Selected papers from the CLARIN Annual Conference*, pages 19–37, Aix-en-Provence, France. Linköping University Electronic Press, Linköpings Universitet.
- Chiarcos, C., and Ionov, M. 2019. Linking the TEI: Approaches, Limitations, Use Cases. In *Digital Humanities Conference 2019 (DH2019)*, Utrecht University, July.
- Chiarcos, C., Fäth, C., and Abromeit, F. 2020. Annotation Interoperability for the Post-ISOcat Era. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5668–5677, Marseille, France, May. European Language Resources Association.
- CoreTrustSeal Standards and Certification Board. 2019. CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022, November.
- Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, C., Katrien, J., Tommaso Agnoloni, D., Venturi, G., Pérez, M., de Macedo, L., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Dargis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.
- Ferger, A., and Jettka, D. 2020. Use cases of the ISO standard for Transcription of spoken language in the project INEL. In *Proceedings of the CLARIN Annual Conference 2020*. CLARIN ERIC.
- Fisseni, B., and Schmidt, T. 2020. CLARIN web services for TEI-annotated transcripts of spoken language. Selected Papers from the CLARIN Annual Conference 2019. Leipzig, 30 September–2 October 2019, pages 12–22. Linköping University Electronic Press, Linköping.
- Frick, E. and Schmidt, T. 2020. Using full text indices for querying spoken language data. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, pages 40–46, Marseille, France, May. European Language Resources Association.
- Hinrichs, E., Hinrichs, M., and Zastrow, T. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29.
- ISO/TC 37/SC 4. 2016. Language resource management – Transcription of spoken language. Standard ISO 24624:2016, International Organization for Standardization, Geneva, Switzerland.
- Janssen, M. 2021. A corpus with wavesurfer and TEI: Speech and video in TEITOK. In Ekštejn, K., Pártl, K., and Konopík, M. (eds.), *Text, Speech, and Dialogue*, pages 261–268, Cham. Springer International Publishing.
- Kisler, T., Reichel, U., and Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.
- MacWhinney, B. 2000. *The CHILDES project: Tools for Analyzing Talk, Third edition. Volume I*. Lawrence Erlbaum, Mahwah, NJ u.a., 3rd edition.
- Ochs, E. 1979. Transcription as theory. In Ochs, E., and Schieffelin, B. (eds.), *Developmental pragmatics*, pages 43–72. Academic Press, New York.
- Parisse, C., Etienne, C., and Liégeois, L. 2020. TEICORPO: a conversion tool for spoken language transcription with a pivot file in TEI. *Journal of the Text Encoding Initiative*, 13, May.
- Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., and Seifart, F. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2657–2666, Marseille, France, May. European Language Resources Association.
- Rehbein, J., Schmidt, T., Meyer, B., Watzke, F., and Herkenrath, A. 2004. Handbuch für das computergestützte Transkribieren nach HIAT. *Arbeiten zur Mehrsprachigkeit, Folge B*, 56:1 ff.
- Remus, S., Hedeland, H., Ferger, A., Bührig, K., and Biemann, C. 2019. WebAnno-MM: EXMARaLDA meets WebAnno. In *Selected papers from the CLARIN Annual Conference*, Pisa. Linköping University Electronic Press, Linköpings Universitet.
- Schmidt, T., and Wörner, K. 2014. EXMARaLDA. In Durand, J., Gut, U., and Kristoffersen, G. (eds.), *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

- Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., and Sloetjes, H. 2009. An exchange format for multimodal annotations. In Kipp, M. (ed.), *Multimodal corpora. From models of natural interaction to systems and applications*, Multimodal corpora. From models of natural interaction to systems and applications, pages 207 – 221. Springer, Berlin [u.a.].
- Schmidt, T., Hedeland, H., and Jettka, D. 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, Linköping Universitet.
- Schmidt, T. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1, 06.
- Schmidt, T. 2016. Construction and dissemination of a corpus of spoken interaction - tools and workflows in the folk project. *Journal for language technology and computational linguistics (JLCL)*, 31(1):127 – 154.
- Selting, M., Auer, P., Barden, B., Bergmann, J., Couper-Kuhlen, E., Günthner, S., Meier, C., Quasthoff, U., Schlobinski, P., and Uhmann, S. 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte*, 173:91–122.
- Sloetjes, H. 2014. ELAN: Multimedia annotation application. In Durand, J., Gut, U., and Kristoffersen, G. (eds.), *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.
- TEI Consortium. 2021. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Technical Report 4.3.0, TEI Consortium, August.
- von Prince, K., and Nordhoff, S. 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.
- Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J., Heringa, J., 't Hoen, P., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S., Martone, M., Mons, A., Packer, A., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.
- Wittenburg, P., van Uytvanck, D., Zastrow, T., Straňák, P., Broeder, D., Schiel, F., Boehlke, V., Reichel, U., and Offersgaard, L. 2019. CLARIN B Centre Checklist (CE-2013-0095), Version 7.3.1, 2019-09-30. Technical report, CLARIN ERIC, September.

CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP)

Yuras Hetsevich UIIP of NASB, Minsk, Belarus yuras.hetsevich@gmail.com	Jauheniya Zianouka UIIP of NASB, Minsk, Belarus evgeniakacan@gmail.com	David Latyshevich UIIP of NASB, Minsk, Belarus david.latyshevich@gmail.com
Mikita Suprunchuk Minsk State Linguistic University, Belarus suprunchuk@mail.ru		Valer Varanovich Belarusian State University, Minsk, Belarus gamrat.vvv@gmail.com

Abstract

This paper represents the CLARIN Knowledge Centre for Belarusian text and speech processing (K-BLP) which is based at the Speech synthesis and recognition laboratory, the United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk. The CLARIN Knowledge Centre for Belarusian text and speech processing is part of the CLARIN ERIC, which holds the European ESFRI (European Strategy Forum on Research Infrastructures) certification as a landmark research infrastructure. Services for text and speech processing, which were developed by the Laboratory, are presented in the article.

1 Introduction

Today, computer technologies are developing rapidly. They capture all new areas of life and fields of knowledge, including those related to language and the transfer of knowledge. For the development of machine dictionaries, translators, search engines and databases, text corpora are increasingly being used. The creation of a corpus can be carried out in different ways, methods, and stages. All of them are quite laborious and require knowledge in linguistics and programming. Proofreading and verification of texts are especially time and human resources consuming stages. In the case of parallel corpora, the problem of sentence alignment is added to this. To solve these and similar problems, a lot of work is being done in the Speech synthesis and recognition laboratory of the United Institute of Informatics Problems of the National Academy of Sciences of Belarus (SSRLab laboratory, <https://ssrlab.by>).

The SSRLab laboratory established the K-BLP Centre in 2020. It provides users with knowledge for text, speech and other data processing for Belarusian, Russian, and English. The K-BLP Centre proposes tools for text, speech and other data processing for languages, especially for the Belarusian language. The centre also offers wide-ranging user support, guidelines and instructions for each service and material.

We are committed to widen the access to Belarusian developments in the computational linguistics environment and popularize our tools within the Republic of Belarus and abroad (Figure 1). It is very important to support available tools and promote them to improve and facilitate the access for researchers in humanities and social sciences that contributes to wide-ranging user support, guidelines and instructions for each service. Primary target audience of K-BLP are researchers in humanities and digital humanities with an interest in different aspects of computational linguistics and natural language processing.

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

Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprunchuk and Valer Varanovich 2022. CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP). *Selected papers from the CLARIN Annual Conference 2021*. Ed. by Monica Monachini and Maria Eskevich. Linköping Electronic Conference Proceedings 189, pp. 46–55. DOI: <https://doi.org/10.3384/9789179294441>

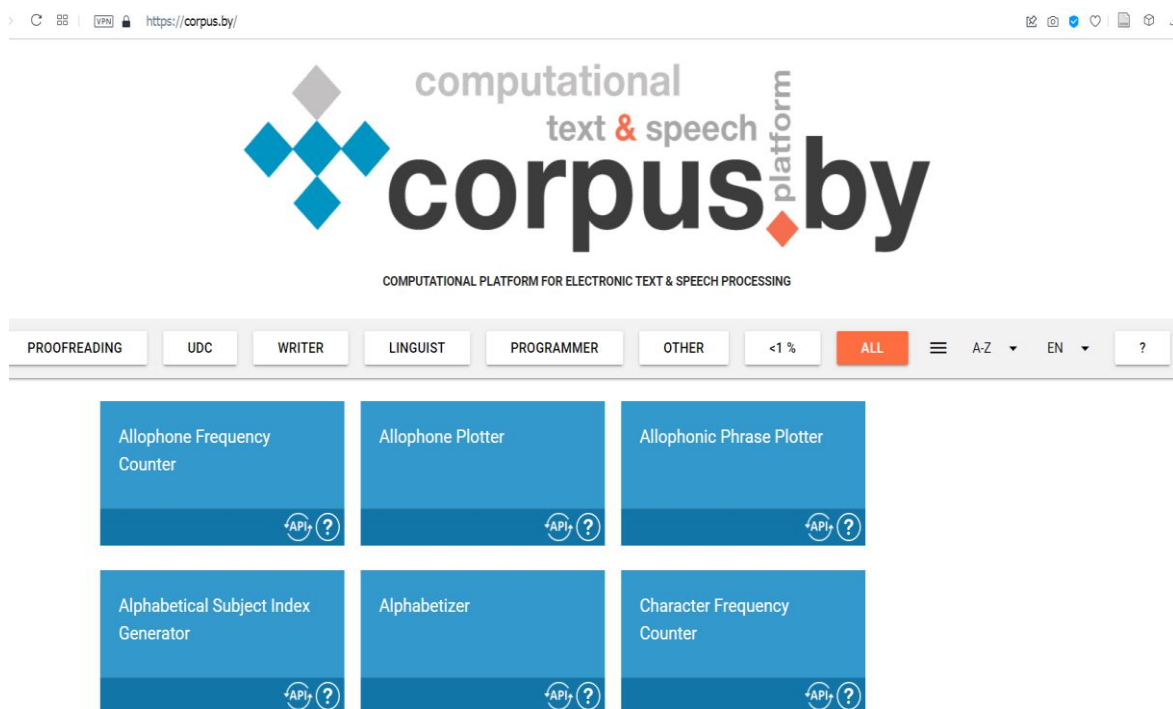


Figure 1. Overview of Belarusian text, speech and other data processors

Next, we will demonstrate a number of services and tools that are used by SSRLab when preparing text corpora. Most of them are developed in the laboratory. Some programs, such as NooJ (Silberztein 2016; NooJ), were created by other people or organizations, but the laboratory offered it as a tool for collecting and processing Belarusian text information. Developing of several services was supported as part of a new CLARIN project in 2021 “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”.

2 K-BLP's Main Aims within CLARIN ERIC Research Infrastructure

The main task of the K-BLP Centre is to extend our resources and tools of natural language processing and organize them according to the data within the CLARIN Resource Families in the examples of other resource families (cf. de Jong, 2020). Increasing the interest in Belarusian developments in computational linguistics and popularizing available tools and resources are the main directions of K-BLP. To follow these aims, we should widen the number of scientific organizations of K-BLP (except the UIIP of NASB), add new resources and structuralize our Belarusian services within the CLARIN classification. It is very important to promote available resources to facilitate access for researchers. That is why we propose wide-ranging user support, guidelines and instructions for each service. We also plan to create and maintain new tools for electronic text and speech processing in the Belarusian language.

At present K-BLP has main strategic priorities such as:

1. To attract other scientific organizations and institutes with research centres for computer processing of the Belarusian language to widen K-BLP (such organizations as Belarusian State University, the Centre for the Belarusian culture, language and literature researches of the National Academy of Sciences and other).
2. To expand K-BLP with such resources as new Belarusian corpora (at least 3), dictionaries (approx. 5-7 items) and other tools for computer processing of Belarusian text and speech information (5-7 tools).
3. To annotate and systematize new resources and tools as consistent with a description of all resources deposited in other CLARIN ERIC centres.

4. To optimize existing resources and tools in K-BLP according to the CLARIN ERIC classification of resources.

5. To organize the overviews of developed Belarusian tools according to the types of data in the resources and listings sorted by language.

6. To provide a user-friendly overview of the available Belarusian language tools in the CLARIN infrastructure for researchers from digital humanities, social sciences and human language technologies.

7. To create and maintain an infrastructure to support the sharing, use and sustainability of Belarusian language data and tools for research in the humanities and social sciences.

We hope to implement our plans listed above in the near future with the help of CLARIN ERIC.

3 K-BLP Centre Initial Activities

The Speech synthesis and recognition laboratory of UIIP NASB established K-BLP Centre in September 2020. Step by step, it started the process of CMDI metadata creation for all online resources. So the part of the services is now available via the VLO. Currently, our centre offers data processing services and tools computational platform for electronic and speech processing platform which includes over 65 services (Dzienisiuk, 2020), a speech intonation analyser and trainer IntonTrainer (Lobanov, 2019), Belarusian NooJ module for convenient processing of Belarusian language via NooJ linguistic development environment), tutorials and exercises. All provided services can also be accessed through the links directly via <http://www.corpus.by/> link. Detailed information is available on the Speech synthesis and recognition laboratory of UIIP NAS Belarus web-site.

The Laboratory works on such main scientific research directions as digitization of cultural heritage, high-quality text-to-speech synthesis, robust recognition of discrete and continuous word sequences, computer systems for the rehabilitation of people with hearing and vision disabilities. In addition, we work with systems, programs and platforms for processing big data, universal algorithms for stationery, online and mobile platforms for asynchronous input and output storing and issuing information from different platforms, semi-automatic systematization and processing of data by administrators of target programs (Figures 2–4).

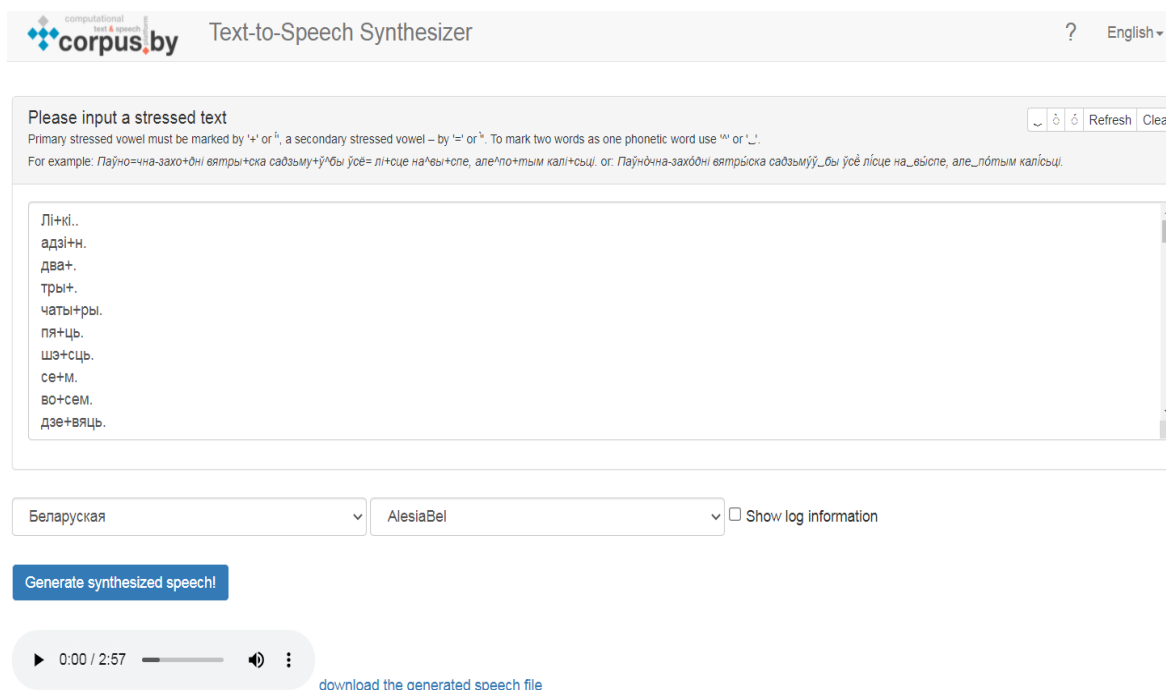


Figure 2. Text-to-Speech Synthesizer

<p>– Data Processing services & tools</p> <ol style="list-style-type: none"> 1. The corpus.by platform with web-based services 60+ for Belarusian and other text, voice and other data processing that can be used by linguists, computational linguists and language engineers, writers and proof-readers. Among others, the platform provides tools for tokenization, morphological analysis, voiced electronic grammatical dictionary, part-of-speech tagging, frequency counter, spell checking and others. 2. IntonTrainer – a speech intonation analyser and trainer, a software system designed to train learners in producing a variety intonation patterns of Belarusian and other speech. 3. Belarusian NooJ module for convenient processing of Belarusian language via NooJ linguistic development environment. 4. Belarusian and other linguistic resources provided for CLARIN Virtual Language Observatory (VLO CLARIN).
<p>+ Tutorials & Exercises</p>
<p>+ Plans</p>

Figure 3. Actual materials of K-BLP

Pitch Plotter

(Part of Speech Synthesis and Recognition Laboratory, [UIIP](#) NAS Belarus)

⊕ The service allows a user to get a graphical image of the pitch frequency contour of a speech phrase online.

[Belarusian](#)
[English](#)
[Russian](#)

🏠 [Landing page for this record at corpus.by](#)

Thematic Lists Collector

(Part of Speech Synthesis and Recognition Laboratory, [UIIP](#) NAS Belarus)

⊕ The service provides lists of words and phrases with at least one vivid example of any allophone or diphone.

[Belarusian](#)
[English](#)
[Russian](#)

🏠 [Landing page for this record at corpus.by](#)

Dialectological Maps

(Part of Speech Synthesis and Recognition Laboratory, [UIIP](#) NAS Belarus)

⊕ The service in an interactive format offers the user information about the dialectological pronunciation of some words in different settlements of Belarus.

[Belarusian](#)
[English](#)
[Russian](#)

🏠 [Landing page for this record at corpus.by](#)

Figure 4. Tools for text processing in K-BLP

Our staff also uses the approaches to process audio and text forms of speech, which are often found in the development of modern systems that work with the input and output of large-scale speech (BigData) on different platforms.

We intend to create and maintain user infrastructure to support the sharing, use and sustainability of big data and tools for research in computational linguistics, the humanities and social sciences. Almost all our digital resources are open, free and available to scholars, researchers and scientists from all spheres through single sign-on access.

All products are made to solve the problems of developing algorithms, resources and methods of Internet input and Internet output of speech, saving and systematizing large volumes of speech. The results can be adapted for wide use in applied and practice-oriented research that requires processing large amounts of data at different levels.

One more task is to provide a user-friendly overview of the available tools for researchers as well as to organize the overviews of developed methods and algorithms according to the types of data in the resources and listings sorted by language. Our team has considerable experience in accumulating big data in different formats and platforms. There are specialists in programming, front- and back-end development, project managers, computational linguists and philologists. We are open to create and develop new resources, tools, algorithms and methods according to users' demands.

Certainly, the K-BLP Help Desk was organised to provide people from Belarus and foreign countries with information about CLARIN ERIC, about Belarusian language in general and computer tools for text researches. There are two main possibilities to apply for such information: via contact page on the web-site <https://clarin-belarus.corpus.by/contacts/> or via contact email which is available on the platform corpus.by. We receive one or two inquires every month.

Besides, several employees who works at SSRLab teaches computer linguistics and other courses (“Problems of AI”, “Computer technologies in linguistics”, “Speech synthesis and recognition”, “Automatic translation”, etc.) at Belarusian State University and Minsk State Linguistic University. Therefore, students (about 90 per year) ask a lot of questions during classes and while preparing their home works and projects. The most frequent questions are the following: How to get acquainted with computer linguistics, esp. with Belarusian computer linguistics? Where one can find digital resources on the Belarusian language and literature? Where one can learn about history of Belarus and the Belarusian language, about traditions and culture?

We try to respond by ourselves, there is a special collection with most useful resources on the web-site about Clarin <https://clarin-belarus.corpus.by/materials/>. A lot of services are presented on the platform www.corpus.by and on the site of our colleagues <https://bnkorporus.info/index.en.html>. There is a nice cooperation between SSRLab and the Institute of linguistics named after Yakub Kolas of the Belarusian Academy of Sciences <http://iml.basnet.by/>, so their consultations are possible, too.

4 Optimization of Information Pre-Processing for the Corpora Creation

Various types of the text processing are important directions of the Belarusian CLARIN Knowledge Centre activities: speech to text and text to speech conversion, spell checking, transliteration, transcription, etc. One of them is NooJ platform. The NooJ application is a shell for word processing and a convenient tool to compile a corpus of texts. It was developed by Max Silberztein (Silberztein 2016), professor of Université de Franche-Comté, France. (Cf. section 5 below.)

At different stages of the corpus preparation, we had specific tasks. To solve them, as well as for other purposes, the laboratory staff developed a number of useful tools and services. Some resources are currently being improved as part of the 2021–2022 project “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”. The following discussion will overview some of them.

When data is collected in large quantities, there may be texts or their fragments in different languages. In our case, it was important to select texts in one language, either only in Russian or only in Belarusian. To check that the text is written in the required language, it is useful to use the service (LanguageIdentifier).

The “Language Identifier” service was developed to the identify the language of the text which has been submitted to the input. For now, the service recognizes five languages: Belarusian, Russian, Ukrainian, English and German. The text language is identified by the service using the statistical method and the rule application method. The priority of “statistics over rules” or “rules over statistics” is determined by the position of a special toggle switch variable. The ability to change the position of this toggle switch is currently hidden from the user. However, if necessary, it can always be used by the developer. The sensitivity threshold of the algorithm, the minimum and maximum number of characters of the text to be processed can be just as easily changed. The plans for the improvement of the service include the ability to define several languages of multilingual text and the generation of statistics on the use of each individual language, the expansion of the language palette, using new identification rules. To access the “Language Identifier” service via the API, one needs to send an AJAX request of the POST type to the address <https://corpus.by/LanguageIdentifier/api.php>.

High quality of the created program largely depends on the source data. It is important to ensure that the collected texts of a compiled corpus, do not contain errors, typos, repetitions, or unnecessary information. One spell checker package for MS Office Word, LibreOffice, OpenOffice, Thunderbird and several web-browsers was developed by one of our colleagues (Praverka pravapisu). It corrects texts fairly well, but it requires special search, download and installation.

In contrast, in our projects the majority of texts are proofread by editors and proofreaders – members of the project team. In addition, all texts are automatically checked by a special free online service which was also developed by the laboratory staff (SpellChecker).

The service receives an electronic document that requires verification. By pressing the “Check it!” button, the service compares text words with words in attached dictionaries. The service qualifies the words of the input text found in at least one dictionary as spelled correctly and discards them. Words that are not found in dictionaries are qualified by the service as misspelled. The service displays them in a list in alphabetical order. Currently, the quality of the text proofreading is an integral requirement for many fields of activity, especially for communication between people and institutions. In addition, spelling-correct electronic text is necessary for proper functioning of computer systems of human-machine communications. The relevance of the service development is also determined by complicated access to processing tools for Belarusian-language texts. The proofreading of an electronic text by machine tools always remains relevant, since manual checking of texts by the user almost definitely means skipping mistakes.

The named service checks Russian and Belarusian documents. To check the Russian language, a well-known dictionary by Andrei Zaliznyak (Zaliznyak 2003) is used. To check the Belarusian language several large modern dictionaries are used, cf. the full list on the web-site (Spell Checker). In addition, the laboratory replenishes its own dictionary, where words that are not included in published editions are indicated because they are recent or used in narrow areas. Some of the mentioned dictionaries are being constantly enlarged.

Among several Belarusian services of spell checking, the “Spell Checker” service was created as one of the stages of preliminary text processing and normalization for a speech synthesizer. It is worth noting that this service covers the orthographic section of the spelling, but not grammar, syntax or punctuation. The correctness of word matching and punctuation is outside the competence of the service and remains for the user or other services that are also involved in the methodology of large electronic texts proofreading using the platform www.corpus.by services. “Spell Checker” service can process both small and large texts. For example, it successfully checked the spelling of legislative codes and literary works with a volume of about 470 000 characters with spaces.

It is important to mention another spell checking tool. There is a specific alteration in Belarusian orthography. The letter *y* and the sound [u] are used after consonants and punctuation marks, and after vowels the letter *ŷ* and the sound [w] are used instead (so called “non-syllable w” or “short w”). Besides this, the sound [w] and the letter *ŷ* alternate with the letters *в*, *н* and sounds [v] and [l] depending on the place in the word and its origin. This alternation has certain peculiarities and limitations, so it was decided not to embed the control of this phenomenon into the general spell checking service, but to develop a separate tool – “ShortUSpellChecker” (Figure 5).

While searching for possible errors, the service not only determines whether the vowel or consonant is before “u/w”, but also analyzes characters that are not letters, if the letter “u” is at the beginning of a word. These characters directly influence the writing of a word. Not all words of the Belarusian language adhere to the general rules for writing the letter “ŷ”. For this reason, the service provides the opportunity to use an exceptions dictionary (can be attached by a special box) or a user list of exceptions. The service processes a text, considering these sets of words. There are special rules for writing abbreviations with the letter “y” in the Belarusian language. To obtain accurate results (since the service does not distinguish abbreviations from other words automatically), the user is prompted to enter the abbreviations that appear in the text in the corresponding field. The service considers the following characters as punctuation marks: “,”, “.”, “:”, “;”, “!”, “?”, “_”, “—”, “(”, “)”. Symbols “[”, “]”, “{”, “}”, “_”, “%”, “№”, “#”, “^”, “\$”, “@” and others are not punctuation marks for the processing algorithm of the service. A hyphen (“-”) is a punctuation mark (identified with a dash) only if it is surrounded by spaces on both sides.

Perhaps, here should be «ў» or «ў»:

There was a letter	Comment
«а у»: ...Мама у трауры...	(«у» after the vowel «а» without a punctuation mark)
«А у»: ...А у іх ёсць пчолы...	(«у» after the vowel «А» without a punctuation mark)
«а у»: ...«Рама» у краме...	(«у» after the vowel «а» without a punctuation mark)
«а-у»: ...На ўкраіне паўднёва-усходні вецер...	(«у» after the vowel «а» and a hyphen)
«ау»: ...Сястра ёсць аусянку...	(«у» after the vowel «а»)
« У»: ...ЛЮДЗІ УСІХ КРАІН, СЯБРУЙЦЕ!...	(«У» after the vowel « » without a punctuation mark)

Perhaps, here should be «У» or «у»:

There was a letter	Comment
«т ў»: ...Кот ў ботах...	(«ў» after the consonant «т» without a punctuation mark)
«т» ў»: ...«Брат» ў космасе...	(«ў» after the consonant «т» without a punctuation mark)
«Ў»: ...На ўкраіне паўднёва-усходні вецер...	(CAPITAL «Ў» IS ONLY ALLOWED IN A TEXT WHERE ALL WORDS ARE WRITTEN IN CAPITAL LETTERS)
«м-ў»: ...Усім-ўсім пра ўсё распавязем!	(«ў» after the consonant «м» and a hyphen)
«бў»: ...Тата любіць бульбу...	(«ў» after the consonant «б»)

Figure 5. Non-syllable U Spell Checker: [u] or [w]

One more interesting service which is used to prepare data for the corpus is “Grammatical Dictionary Processor” (Grammatical). This service allows the user to receive previously loaded and converted to the required format lexicographic data of the grammar dictionary in the form of the HTML table, and to receive SQL instructions for creating a database that contains the entered information in a structured form.

Many text analysis-oriented systems need extensive and well-structured vocabulary databases – for example, automatic annotation and abstracting systems, systems of market analysis, legal linguistic examination. In addition, the vocabulary base can become the basis of commercial products – such as programs designed to help the user improve the grammar of the text he or she wrote, or popular entertainment applications that offer word games to the user. Filling such vocabulary databases (and especially filling grammatical dictionaries) is a very time-consuming and painstaking process. “Grammatical Dictionary Processor” service is designed to simplify and automate this process in the case of working with Belarusian-language data. Thus, the service devotes to provide additional support to strengthen the position of the Belarusian language in the electronic space.

The results of the “Grammatical Dictionary Processor” service were repeatedly applied in other tools of the Corpus.by platform.

The service processes texts only in Belarusian. It is available via the API too. The details are presented here: <https://ssrlab.by/en/8071>.

This tool is under development yet in a frame of the new project for 2022 year “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”.

It is planned that part Belarusian general corpus will comprise oral speech, recordings of speeches and spontaneous conversations. The “Thematic Speech Recognizer” program <https://ssrlab.by/en/4962> was developed to speed up the decryption and transcription of audio files. It allows the user to convert speech to electronic text online. A phonogram no larger than 20 MB is given at the input to the service. It provides a recognized electronic text of the phonogram at the output. The soundtrack can be selected from the provided examples, downloaded to the service from the computer's hard drive in WAV format, and can also be recorded online.

Speech recognition has great scientific perspectives and wide possibilities of application in many “human-machine” systems, which are built on the basis of speech communication. There are other

areas that are particularly in need of speech recognition services. For example, journalism, shorthand and many others. In particular, the recognition of Belarusian speech, which becomes possible with the help of this service, will allow the full development of Belarusian technical sciences, including robotics. At the moment, the service is a demonstration that recognizes the Belarusian language of the following thematic domains: clothes, cities, numbers, etc. The list of domains will be continued. The tool is implemented and works according to the instruction on creation of programs on the basis of CMU Sphinx (CMUSphinx). It is available via the API.

When texts are selected, it is often necessary to get their quantitative characteristics. For example, it is usually useful to make a list of wordforms used. The “Tokenizer” tool was created for this and similar tasks. The token is a wordform, e. g. *come, came, coming*, or *cat, cats*, or Belarusian cases of ‘hand’ *рука, руку, руку, руку*, etc. “Tokenizer” is intended to locate tokens in the text that requires tokenization. It is sent to the service input. After its processing, the user receives a list of the extended tokens on the output. Figures and punctuation marks are processed too. The service handles Belarusian, Russian and English. It is available via the API. The details are presented here: <https://ssrlab.by/en/5900>. The tool is also being developed as part of the “Preparation of available K-BLP tools and resources for the metadata ingestion into CLARIN virtual language observatory (VLO) and representation in CLARIN Resource Families”.

5 Work on Legal Texts

The Speech synthesis and recognition laboratory is working to create a parallel body of legal texts (Hetseвич 2021). The creation of a corpus of legal texts will allow solving several important tasks to improve the automatic processing of texts in Belarusian. First, the corpus will allow us to conduct comparative research and identify features of the Belarusian language in comparison with Slavic and other European languages. Second, it is possible to create various dictionaries, both monolingual and multilingual, as well as a linguistic knowledge base for the machine translation system. The corpus can also be a basis for creating a variety of morphological and syntactic grammars that can be used for automatic grammatical analysis of texts in the Belarusian language. Finally, with the help of such a corpus, it is possible to identify the features of the legal style in Belarusian.

For the created corpus the codes of the Republic of Belarus – the most important (together with the Constitution) legislative acts regulating civil legal relations in various spheres of public activity – were taken as a base. It is planned to present the texts of the codes in two official languages of the Republic of Belarus – Belarusian and Russian, and in the future – in other languages. At the beginning of 2022, 18 out of 26 codes were processed and uploaded to the project page, the total number of word forms in the existing building is about 1 million. The results of the work are at <https://ssrlab.by/7804>.

The corpus of codes was also compiled in the NooJ format (NooJ), and a trilingual Belarusian-Russian-English dictionary of legal terms was created on its basis (Figure 6).

Yuras Hetseвич and Sviatlana Hetseвич from SSRLab laboratory and Yauheniya Yakubovich from Universitat Autònoma de Barcelona created a module to gather Belarusian texts and process them on the basis of NooJ (Hetseвич Y. and Hetseвич S., 2012). The program helps to ‘develop linguistic resources to formalize various linguistic phenomena at the orthographical, lexical, morphological, syntactic and semantic levels, for any natural language’ (NooJ). In addition, with its help everyone can form their own text corpus, select concordances for the analysed words, search in accordance with different language parameters and get the necessary statistical information about certain linguistic facts.

The Budgetary Code, Water Code, Electoral Code, Civil Code, Housing Code, Tax Code, Marriage and Family Code, Forest Code, etc. have been translated till now. When 17 law codes were translated into Belarusian, a group of specialists compiled a unified corpus in NooJ format to create a dictionary of legal terms and expressions. The total number of word tokens is 1,043,018 and 731,584 word forms.

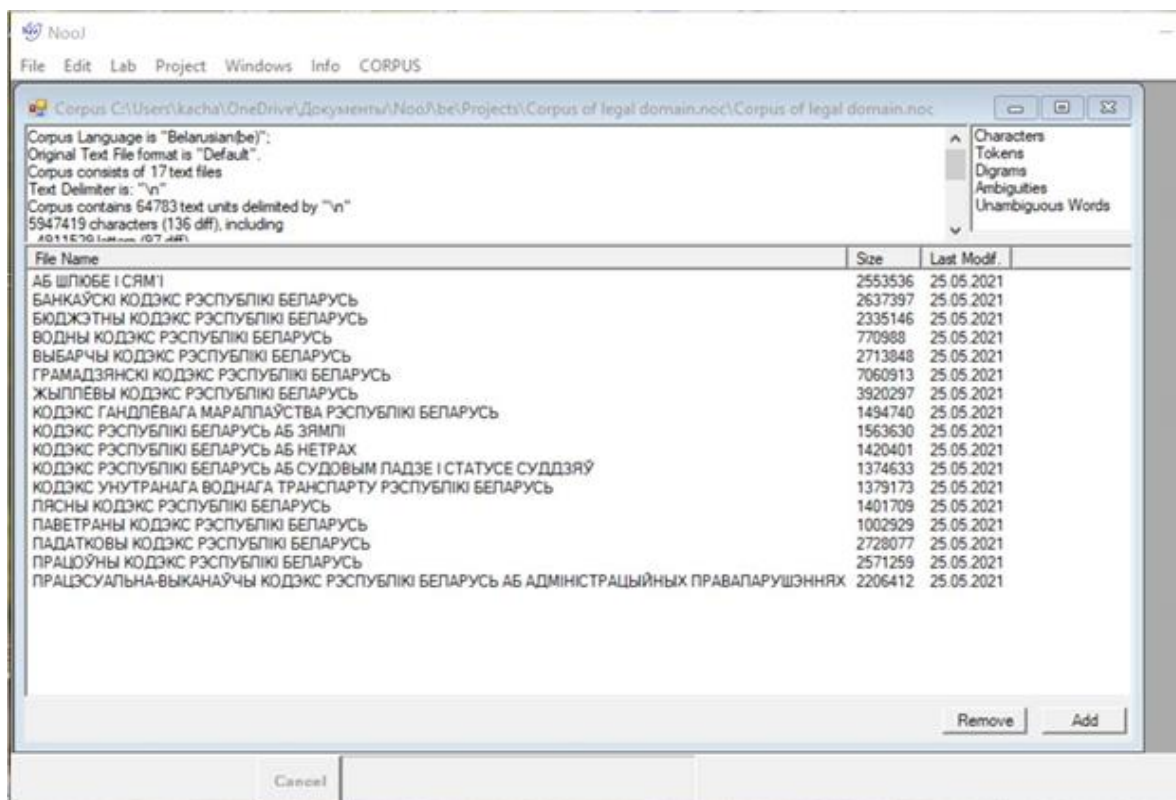


Figure 6. Corpus of translations of legal texts prepared in NooJ

As the Belarusian NooJ module already had the general vocabulary "general_be.dic" it was decided to create an additional dictionary for general unknown words (nearly 150 word forms in initial form) and a law dictionary (nearly 200 forms). Every initial form (lemma) from the list of unknown general words was assigned a morphological class. This class shows the flexion features, i. e. how the word changes. Then the law dictionary in the Belarusian was compiled for NooJ format and now is available for further text processing with Belarusian texts of any domains. It is an addition to the main NooJ dictionary for the Belarusian language.

All terms of the Belarusian law dictionary Law_codes_be.dic were correlated with their Russian equivalents from the parallel corpus of a legal domain. The next step was their translation into English.

Next, several grammars were developed that show the possibilities of applying the established legal corpus in solving various problems of machine texts processing.

6 Conclusion

Building and running a distributed knowledge centre K-BLP for computational linguistics and natural language processing requires samples, text descriptions, demos, courses and possible contacts with specialists of natural language approaches of Belarusian.

K-BLP provides knowledge about tokenization, morphological analysis, voiced electronic grammatical dictionaries, part-of-speech tagging, frequency counting, spell checking, text classification and other tools, algorithms and methods used in speech and text processing. It offers special courses in language processing, data analysis and collecting research data for the fast entrance of humanities and others into the digital world of Belarusian data processing.

The Speech synthesis and recognition laboratory organises several courses in universities to educate students and researchers in computer linguistics. Several online education materials in English were prepared, such as "Lab 0 – How to be acquainted with text and speech processing services in 10 days?". Introduction into the CLARIN project will be presented here, too. All this will allow the representation of different tools for computational processing of Belarusian for all interested in it including foreign scientists and partners.

We are aimed at collecting Belarusian-language linguistic and computer resources for manual and automatic processing in one unit for popularizing the Belarusian language as much as possible. There

is a variety of developments in Belarusian, but they are not in the public domain. For this, we want to conduct research in computational linguistics and modern standard Belarusian and represent results within the K-BLP Centre. The future idea is to participate with other CLARIN centres in joint European projects. The plan is to prepare main services and tools from Computational platform for electronic text & speech processing www.corpus.by for CLARIN Virtual Language Observatory.

References

- Silberztein, M. 2016. *Formalizing Natural Languages: The NooJ Approach*. Wiley Eds. Hoboken, NJ, USA. 346 pp.
- NooJ: A Corpus Processor. URL: <https://www.nooj-association.org/>.
- Jong de, J., Maegaard, B., Fišer, D. [et al.] 2020. Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN. *Proceedings LREC 2020*, 12th International Conference on Language Resources and Evaluation, ELRA. URL: <https://www.aclweb.org/anthology/2020.lrec-1.417>.
- Dzienisiuk, D. A., Zianouka, Ja. S., Drahun A. Je. [et al.]. 2020. Platforma dlja apracouki tekstavaj i hukavoj infarmacyi dlja roznych tematycznych damienau bielaruskaj movy. *Jazykovaya lichnost' i ěffektivnaya kommunikatsiya v sovremenom polikul'turnom mire: materialy VI Mezhdunar. nauch.-prakt. konf., posvyashch. 100-letiyu Belarus. gos. un-ta, Minsk, 29–30 okt. 2020 g.* / Belarus. gos. un-t ; redkol.: S. V. Vorobyeva (gl. red.) [i dr.]. Minsk, BGU: 69–74.
- Lobanov, B. and Zhitko, V 2019. Software Subsystem Analysis of Prosodic Signs of Emotional Intonation. *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings* / eds. Albert Ali Salah, A. Karpov, R. Potapova. Springer: 280–288.
- LanguageIdentifier. URL: <https://corpus.by/LanguageIdentifier/?lang=en>.
- Praverka pravapisu. Праверка правапісу // Беларускі N-корпус. URL: <https://bnkorporus.info/spell.html>.
- Spell Checker. URL: <https://corpus.by/SpellChecker/?lang=en>.
- Zaliznyak, A. A. 2003. Зализняк, А. А. Грамматический словарь русского языка: Словоизменение [Grammaticheskij slovar' russkogo yazyka: Slovoizmenenie]. Москва, Рус. словари [Moscow, Rus. slovari]. 800 pp.
- ShortUSpellChecker. URL: <https://corpus.by/ShortUSpellChecker/?lang=en>.
- GrammaticalDictionaryProcessor. URL: <https://corpus.by/GrammaticalDictionaryProcessor/?lang=en>.
- CMUSphinx. URL: <https://cmusphinx.github.io/wiki/about/>.
- Hetsevich, Yu. 2021. Creation of a legal domain corpus for the Belarusian NooJ module: texts, dictionaries, grammars / Yu. Hetsevich, Ya. Zianouka, V. Varanovich, M. Suprunchuk, Ts. Prakapenka, Dm. Dzenisiuk. *15th International Conference NooJ 2021: book of abstracts / Virtual conference* ; ed. M. Bigey [et al.]. Besançon, France: 36-37.
- Hetsevich, Y. and Hetsevich, S. 2012. Overview of Belarusian and Russian dictionaries and their adaptation for NooJ. *Automatic Processing of Various Levels of Linguistic Phenomena: selected papers from the NooJ 2011 Intern. conf.* / eds. K. Vučković, B. Bekavac, M. Silberztein. Newcastle, Cambridge Scholars Publishing: 29–40.

Curation Criteria for Multimodal and Multilingual Data: a Mixed Study within the QUEST Project

Amy Isard
IFUU/IDGS

University of Hamburg, Germany
amy.isard@uni-hamburg.de

Elena Arestau
IFUU

University of Hamburg, Germany
elena.arestau@uni-hamburg.de

Abstract

We conducted a user survey and expert interviews within the ongoing QUEST project to get an impression of the needs of users and researchers who are working with multimodal and multilingual linguistic corpora. This contribution describes the design and results of the mixed study, whose main goal is to improve the reuse potential of these resources, and to identify concrete topics which are important for the curation of such data.

1 Introduction

Existing approaches to manually or automatically measuring data quality are mostly generically based and aim at the evaluation of research data in general. They do not provide detailed guidance on research data management for specific resource types but simply reference the standards of a community without specifying them further. The research described in this paper was conducted during the ongoing QUEST project¹ (Arkhangelskiy et al., 2021; Arestau, 2021; Hedeland, 2022), which has the aim of enhancing research data quality and re-use for audiovisual annotated language data, and improving adherence to the FAIR principles (Wilkinson et al., 2016).

QUEST develops discipline-specific curation criteria that are tailored to specific re-use scenarios. With regard to concrete re-use scenarios for research data from the fields of language documentation, multilingualism research, sign language and oral history, we define requirements for data, their structure and content. The studies reported here relate specifically to the project's work on curation criteria for multimodal data and for the linguistic secondary use of multilingual data. We set out to get an impression of the needs of corpus researchers, and the obstacles which they currently encounter in re-using or creating such data. To evaluate the reuse potential of such language data, we are developing technical and documentary standards for the various relevant resource types and their metadata alongside discipline-specific curation criteria geared to specific reuse scenarios. Based on this, we have identified concrete topics which are important for the curation of such data, and they have informed our development of the tools and knowledge-base in the QUEST portal.

The rest of this paper is structured as follows. In Section 2 we first give more background and details about the QUEST project. We then define what we include as multimodal and multilingual corpora. We present the design of the survey and interviews in Section 3, followed by their results in Section 4. Finally in Section 5 we discuss how the results and outcomes have influenced our work on the QUEST project.

2 The QUEST Project

The full title of the QUEST project is “Quest: Quality - Established: Testing and application of curation criteria and quality standards for audiovisual annotated language data”. The project is funded by the German Federal Ministry of Education and Research (BMBF) and is one of twelve projects in different disciplines which all aim to improve the re-use potential of scientific research data. Although the project is based in Germany, its results are intended to be used by the global research community. The QUEST

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>

Amy Isard and Elena Arestau 2022. Curation Criteria for Multimodal and Multilingual Data: a Mixed Study within the QUEST Project. *Selected papers from the CLARIN Annual Conference 2021*. Ed. by Monica Monachini and Maria Eskevich. Linköping Electronic Conference Proceedings 189, pp. 56–67. DOI: <https://doi.org/10.3384/9789179294441>

project is based around seven research centres which were already part of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)² (Hedeland et al., 2018): the Data Center for the Humanities (DCH)³ and the Department of Linguistics (IfL)⁴ at the University of Cologne, the Endangered Languages Archive (ELAR)⁵, formerly at the SOAS University of London and since 2021 at the Berlin-Brandenburg Academy of Sciences and Humanities, the World Languages Institute⁶ at the SOAS University of London, the Hamburg Centre for Language Corpora (HZSK)⁷ and the long-term project INEL⁸ at the University of Hamburg, and the Leibniz-Centre General Linguistics (ZAS)⁹ in Berlin. For the QUEST project, the CKLD centres were joined by two further partners who brought expertise in different research areas: the Institute of German Sign Language and Communication of the Deaf (IDGS)¹⁰ at the University of Hamburg, and the Archive for Spoken German (AGD)¹¹ at the Institute for German Language (IDS) in Mannheim. More details on the background to the project can be found in Arkhangelskiy et al. (2021).

QUEST has designed an evaluation system which combines several approaches to assessing data quality. A data review process is provided which consists of guided online surveys, web-based quality checks and subject-specific reviewing. The QUEST portal will cater both to users who are in the process of designing and creating a corpus and those who have already completed their corpus collection and possibly its annotation. For the former it will provide a knowledge base and walk-through on various topics, such as annotation schemes, metadata and anonymisation, and providing links to existing resources. For the latter, users who have a completed corpus may want to deposit it in an archive for long-term storage and to make it findable and accessible for re-use by other researchers. The QUEST project does not itself provide storage; archives which choose to make use of the QUEST services can direct corpus creators to the QUEST portal and questionnaires, where their data can be evaluated and a report generated. Automated tools check for conformity against various criteria, for example whether the structure of the corpus conforms to what is specified in the metadata. The archive can then judge whether the corpus fulfils their deposit criteria and inform the corpus creator of any improvements which are needed. Where automatic checks are not possible, further assessment may be carried out by domain experts, in collaboration with the archive.

2.1 Multimodal and Multilingual Corpora

There are many different descriptions of what exactly is meant by “multimodal” and “multilingual” corpora. Below we provide the definitions for both terms as used in the QUEST project.

Allwood (2008, p. 210) discusses many possible meanings for what a multimodal corpus is and settles on “a digitized collection of audio- and video-recorded instances of human communication connected with transcriptions of the talk and/or gestures in the recording”. Foster and Oberlander (2007) state that “A multimodal corpus is an annotated collection of coordinated content on communication channels such as speech, gaze, hand gesture, and body language, and is generally based on recorded human behaviour.” In the QUEST project, we include video or audio corpora of spoken or signed language, which have various levels of annotation including, at a minimum, transcriptions (of spoken language) or translations (of signed language).

Concerning multilingual resources, we take a broad definition for the concept of multilingualism: “(...) ist mehrsprachig, wer sich im Alltag regelmäßig zweier oder mehrerer Sprachvarietäten bedient und auch von der einen in die andere wechseln kann, wenn dies die Umstände erforderlich machen (...)” [a multilingual person is someone who regularly uses two or more language varieties in everyday life and can

²<https://ckld.uni-koeln.de>

³<https://dch.phil-fak.uni-koeln.de>

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

⁵<https://www.elararchive.org>

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.idgs.uni-hamburg.de/en.html>

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

also switch from one to another if circumstances make it necessary] (Lüdi, 2011, p. 18). It is not only the number of languages used in the corpus which is relevant, but also the potential multilingual background of the participants, which “enable linguists to carry out analyses about multilingual individuals, multilingual societies or multilingual communication” (Schmidt and Wörner, 2012, Introduction). A review of the literature reveals a broad range of multilingual corpora that focus on different aspects of research (Schmidt and Wörner, 2012; Hedeland et al., 2014). For instance, for contact language corpora the typological distance between languages is relevant, “since this helps to predict the type of interference that may occur” (Thomason, 2010, p. 40). The status of the languages and sociolinguistic factors are also relevant for such resources. For language acquisition corpora several factors must be considered: the individual requirements of the learners, their mother tongue, particularities in the acquisition of language and the attitude towards the learning of language and the specification of the regional variety (Bergmann, 2018, p. 28). We do not for the purposes of this project include corpora that consist of a collection of otherwise monolingual sub-corpora.

3 Study Design and Participants

It was decided that the most effective method for designing this study would be a mixed approach (Rubin and Rubin, 2005) which involves both a quantitative user survey and qualitative interviews with researchers and experts as data providers, users and creators. Both in the survey and in the expert interviews, the participants came from a wide range of research areas. We were interested in researchers involved both in corpus creation and re-use, and indeed there is not a clear boundary between the two, as many survey participants mentioned that they had used an existing corpus but added their own annotations (see Section 4.1).

3.1 Survey Design

The target groups of our survey were researchers who were involved in projects dealing with multimodal or multilingual data. The survey was open between July 2020 and March 2021. During this time it was advertised a number of times via twitter, DhD-blog, corpora-list, linguistlist, internal mailing lists, and professional associations.

For the conceptualisation of the survey we were informed by several studies dealing with the curation, management and reuse of research data (Ferus et al., 2015; Fandrych et al., 2016; Arndt et al., 2018). Based on these studies and on preliminary criteria we developed a catalogue of questions. We conducted a pilot survey with five participants prior to the survey release and then finalised the questions together with other project members.

The survey was created using the LimeSurvey online survey tool¹² and was available in German and English via any web browser. The survey contained a maximum of 74 questions, but was designed so that later questions were presented depending on the answers to earlier ones, to avoid participants having to see and respond to questions which were not relevant for them. Data from the survey were handled anonymously, to ensure that there would not be any privacy concerns and that participants would feel free to make negative comments if necessary.

Every survey participant was asked to choose one corpus they would like to discuss. The questionnaire consisted of seven subject blocks covering the following topics relating to that corpus. The questionnaire subjects were chosen based on the FAIR principles and the objectives of the QUEST project. In all cases, questions which might lead to a loss of anonymity, such as the name of the corpus, were optional. Some questions had multiple choice answers, and others allowed free text input. At the end of each section, there was a text field where participants could add any extra comments. The questionnaire blocks were as follows:

1. **Corpus General Information** - which format the corpus was in, which primary data (video and/or audio) it contained, what questions the participant was researching.
2. **Languages** - the languages present in the corpus, including primary data and translations.

¹²<http://www.limesurvey.org>

3. **Transcription and Annotation** - which transcriptions and annotations were already present, and which (if any) were added by the participant.
4. **Anonymisation** - what type of anonymisation was present, if any, and whether it was noticeable to the researcher, or affected their research.
5. **Metadata** - which metadata and/or bias statements were included in the corpus, if any, and whether they were considered to be sufficient.
6. **Access** - How the participant accessed and worked with the corpus, and any problems which they encountered.
7. **Participant General Information** - the country, type of institution and research area the participant works in.

These subject blocks were chosen so that we could obtain information about the corpora described, including languages, annotations, and metadata and also about how the corpus was used by the researcher, and any problems and barriers to re-use which they encountered.

3.2 Survey Participants

The survey was fully completed by 44 participants, and we include only completed results in our analysis. Although this number of responses does not allow us to draw firm quantitative conclusions, we were able to observe some trends and received useful feedback in the free-form comment fields.

We had attempted to find a balance, keeping the survey short enough to encourage researchers to participate but with enough questions to provide us with the necessary information. Most participants who answered any questions did go on to complete the full survey, so we do not think that the length contributed to the low response rate. We had intended to publicise our survey at the conferences and workshops we attended in 2020 and 2021 but the global pandemic meant that these all took place online. The online platforms which were used could not provide a good virtual substitute for the serendipitous interactions with other attendees which typically occur during coffee breaks, and where we could have promoted our survey informally.

The number of questions answered by each participant ranged between 23 and 53, with an average of 35. The participants currently work in 13 different countries: Germany, Italy, Australia, France, Brazil, Ireland, USA, Hungary, Canada, Czech Republic, UK, Tunisia and Austria, with the majority in Germany (62%). They are active in a wide range of research areas: Linguistics, Corpus Linguistics, Computational Linguistics, Historical Linguistics, Multilingualism, Language Acquisition, Sociolinguistics, Translation and Interpreting, Computer Science, Virtual Agents, Multimodal Behaviour, Finance and Sociology. They are employed by universities (72%), data centres, companies and archives.

3.3 Interview Design and Participants

The two authors of this paper carried out qualitative semi-structured interviews to gather deeper insights into the experiences and needs of the experts as data providers and users. We conducted 20 interviews with experts in the areas of multilingual and/or multimodal corpora, and each interview lasted between 45 and 60 minutes. The interview topics were based on the survey, and each interview consisted of three key sessions: 1) Transcription and Annotation, 2) Formats, Standards and Metadata, and 3) Obstacles, Wishes, Suggestions and Challenges.

At the beginning of each interview, topics related to the three key sessions were presented to the experts:

1. Concerning the transcription and annotation of multilingual or multimodal corpora, what are the best practices and tools in your research community? Are there commonly accepted format standards? What conventions do you use?
2. What are the best ways to anonymise the data?

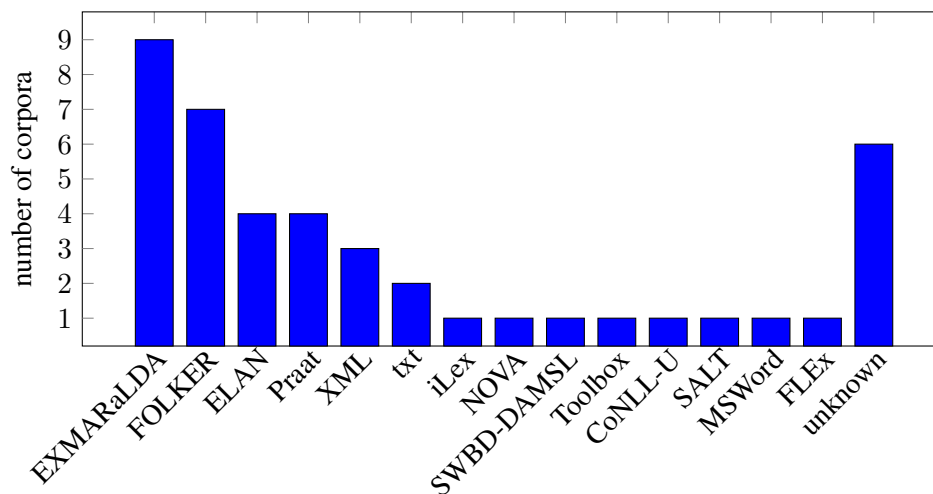


Figure 1: The Tools and Formats of the Corpora from the Survey

3. What are the areas in particular need of investigation in the area of multimodal and multilingual corpus research?

The experts were free to express themselves on other subjects and were not restricted to the suggested topics.

The 20 experts were chosen to represent a wide variety of research interests within the subject areas of multimodal and multilingual corpora. The experts worked in universities and research centres in Germany, UK, Australia, Denmark, Ireland, USA, Norway and Italy, and their main areas of research included:

- documentation of endangered languages
- semiotics of multimodal signed and spoken language interaction
- multi-party interaction
- non-verbal communication and the socio-linguistic contexts of communication
- sign language corpora
- the interface between spoken language and gestural behaviour
- interpreter-mediated interaction within the study of community interpreting
- the analysis of learner languages and errors
- second language acquisition and first language attrition
- contrastive research

4 Survey and Interview Results

In this section we will present the results of the survey and expert opinions on the various different topics described in the previous section. In each subsection we first report some findings from the survey and then summarize the related opinions from the expert interviews.

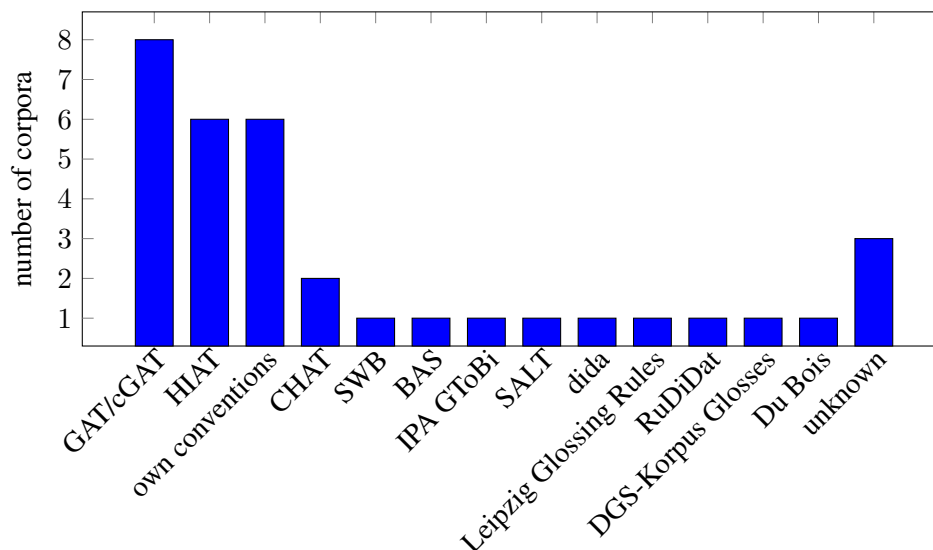


Figure 2: Transcription Conventions from the Survey

4.1 Transcription, Translation and Annotation

The survey revealed that corpora described by the participants contained more than 30 different languages as primary data. Audio recordings were present in 84% of the corpora and video in 41%. Translations were present in 36% of the corpora, and 30% of participants stated that their research questions were in the area of multilinguality. This variety of corpora were reflected in a large number of different formats. We provided the names of common tools but also allowed respondents to add their own. The results can be seen in Figure 1. Several of these answers (XML, txt, MSWord) give no information about the format of the corpora, and a number of researchers were unaware of the corpus format, probably because they had accessed it via a web browser.

Despite the relatively small number of respondents in the survey, a range of transcription conventions were used, as shown in Figure 2. Sixty-seven percent of the corpora reported in the survey were already annotated, and 47% of the respondents added further annotations of their own. Some annotations were included in the majority of the corpora, such as part of speech tagging and lemmatisation, but there was also a long tail of annotations which appeared only once in our survey, as can be seen in Figure 3.

For some sub-areas there was a more consistent picture; for instance, many of the survey respondents who stated that their research was in the area of multilinguality used the editors EXMARaLDA¹³ (Schmidt and Wörner, 2014) and ELAN¹⁴ (Wittenburg et al., 2006) and the transcription conventions CHAT (MacWhinney, 2000) or HIAT¹⁵ (Rehbein et al., 2004), and the same tools and conventions were among those frequently mentioned by the experts in this area.

Several experts in the multimodal domain remarked that it was not always possible to train annotators in the use of tools, because of the time required, and therefore annotation was done in simple text files or spreadsheets. In one case an expert said explicitly that they had decided that they had calculated the trade-off between time taken to train annotators and time spent correcting mistakes in spreadsheets and decided that the latter was less expensive. They also said that it was not possible to rely on the continued availability of specialized annotation tools, whereas commercial spreadsheet software was likely to remain largely unchanged for many years.

Two challenges in particular were mentioned by the experts concerning translations in the field of multilinguality. Firstly the four eyes principle should be used, so that at least two translators have read each text, and secondly there is a need for high language competence, both for the target language and for the source language in the corpus. It is important to have native speakers for translations, so that you

¹³<https://exmaralda.org/en>

¹⁴<https://archive.mpi.nl/tla/elan>

¹⁵https://www.exmaralda.org/pdf/HIAT_EN.pdf

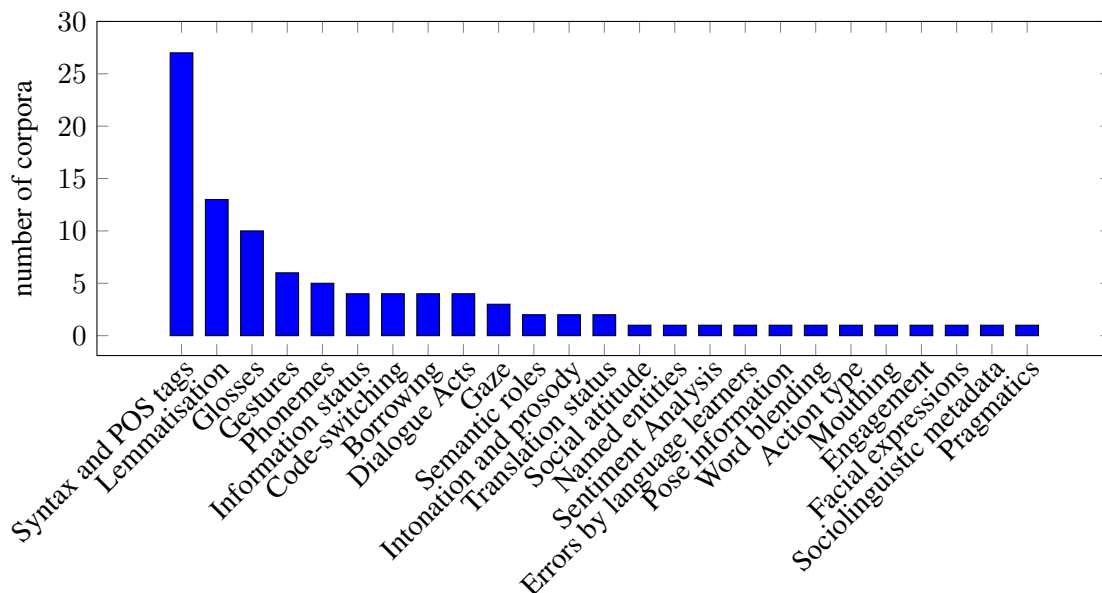


Figure 3: Annotation Types from the Survey

can trust the translation and use it for further research. The four eyes principle can then help in coming to a consensus on problematic cases.

4.2 Anonymisation and Pseudonymisation

We asked the survey participants various questions about the anonymisation of the corpus which they worked with. Fifty-two percent said the the corpus had been anonymised in some way, and the details are shown in Figure 4. Where audio recordings were present, 55% of survey participants said that they had been anonymised, but this was only the case for 17% of video recordings. Of the corpora that were anonymised, transcriptions had been anonymised in 86%, translations in 71%, and annotations in 62%. Audio was anonymised most often with white/brown noise over the affected areas (63%) or with silence. Video was anonymised either by blurring affected areas or by adding black shapes. Translations, transcriptions and annotations were anonymised in a number of ways:

- Entity name pseudonymisation (e.g. Kiran replaced with Anita)
- Enumerated entity name categorisation (e.g. Kiran replaced with PERSON1, Haruki replaced with PERSON2)
- Entity name removal (e.g. Kiran and Lagos both replaced with XXX)
- Entity name categorisation (e.g. Kiran replaced with PERSON, Lagos replaced with PLACE etc)

We also asked the survey participants whether they had noticed the anonymisations (78%) and whether they felt that their work had been affected in any way (13%). Three participants gave details of negative impacts on their research: one said that white noise in the audio stream meant that they were unable to hear the pitch contours of the speech, one that anonymisation of names meant that information about pronoun choices was affected, and another that audio anonymisation prevented them from being able to study rhythmic patterns and long range phenomena in speech.

The experts agreed that informed consent is very important when recording language data of any kind, and that it is important to also protect the anonymity of third parties mentioned by corpus participants, since these people will not have a chance to give their consent. One expert who works with video documentation of small community languages remarked that the choice of whether or not to anonymise a corpus must be based on the wishes of the language community. In their experience, participants were

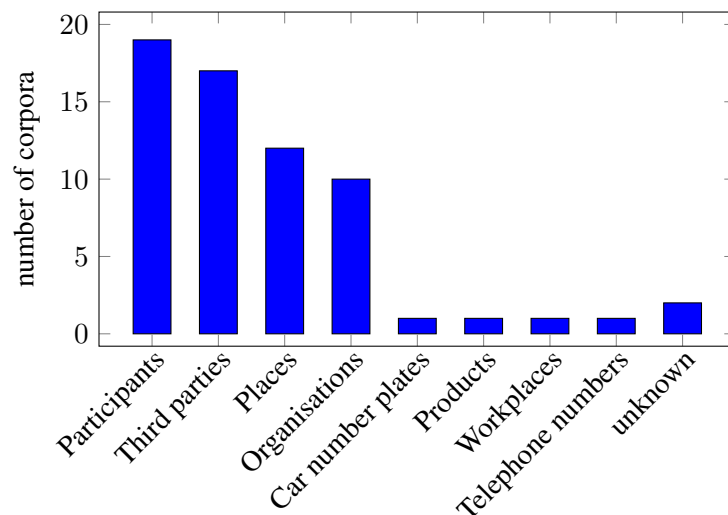


Figure 4: Corpus Anonymisation from the Survey

proud to have taken part and were keen to have their videos available to all. In this case, it was more important to give credit to the participants rather than to anonymise their data. On the other hand, another expert working in sign language research pointed out that in smaller language communities, people are more easily recognisable. For example, when the researcher was giving a presentation on the data, members of the audience recognised some of the participants in the videos shown as examples. In these cases, care must be taken to balance the need for anonymisation against the desire for recognition. The amount of anonymisation necessary can also depend on the licence under which a corpus will be released.

4.3 Metadata

Eighty-eight percent of the survey participants stated that their corpus provided metadata, 7% that it did not, and 5% did not know. Ninety-four percent of those who had metadata stated that it was sufficient for their research needs. Where this was not the case, we asked what was missing, and examples included:

- detailed information of the recording location, the people present, and the position of the recording equipment
- full information about the languages spoken by subjects in a learner corpus

We also asked whether the corpora had documentation of potential biases in the data, for example in the form of a Data Statement¹⁶ (Bender and Friedman, 2018) or Data Sheet (Gebru et al., 2021); this was only confirmed in 7% of cases, while 49% stated that there was none present, and the remainder that they did not know. These statements differ from conventional metadata in that they focus on documenting biases inherent in the data, so as to prevent inaccurate conclusions being reached from overgeneralizations based on a sample from a small subsection of a population. This is particularly important when natural language processing techniques are being carried out on datasets, but it is valuable for any corpus to document the characteristics of the linguistic background not only of the participants in a corpus but also of the annotators and curators.

The experts all emphasised the importance of documenting the process of corpus creation in detail, over and above what is included in the metadata, so that when questions arise later, the answers can be found in the documentation. All experts also stated that detailed metadata for corpora are essential, and should adhere to the standards of the appropriate research community.

¹⁶<http://techpolicylab.uw.edu/data-statements>

4.4 Corpus Findability, Storage and Access

Some of the survey participants had participated in the creation of their corpus and so were not presented with questions about findability and access. Most of the participants who searched for a corpus found it easy to locate and use their chosen corpora (84%). Sixty-five percent had to provide some information before accessing the corpus, such as email address, affiliation, real name, reasons for access, or a signature. Fifty-nine percent had to wait for some form of verification before accessing the corpus. Seventy percent accessed the corpus via a web interface and the rest downloaded it for offline use.

The corpora were used in a variety of ways, including:

- search queries (77%)
- reading transcripts (54%)
- quantitative analysis (48%)
- listening to recordings (45%)
- watching recordings (27%)
- computational models (25%)
- reading translations (23%)

In the free comment section, several issues with corpus access were mentioned multiple times. One participant remarked that it was very difficult to find corpora, since they are not all stored in a single location, and two remarked that they are almost impossible to find through a web search. There are existing solutions to this problem, including the CLARIN Portal¹⁷, but it appears that some respondents were not aware of this resource; the QUEST portal will provide a link to this and other resources in its Knowledge Base.

When asked about barriers to access and reuse, several mentioned the issue of funding - some corpora have expensive licences, and if a university or department does not already have a licence, the funds must be obtained from an individual project or research grant. It was also remarked that even if data was freely available, there remained issues regarding long term availability, since many corpora and software tools vanish over time if there is no structure for their maintenance.

The experts mentioned that for corpus creators it is important to consider where a corpus will be deposited when designing the initial corpus collection study. The experts in multimodal corpora also brought up the issue of funding; where a large corpus is concerned, and particularly if there will be many video files, funding for corpus storage can be an issue, and must be budgeted for in project applications.

5 Conclusions

This study has provided information about the experiences of multimodal and multilingual corpus users and creators, which we have used to inform the design and content of the QUEST project portal. We heard from corpus creators and users from numerous countries and research areas, and were able to create an overall picture of the current needs and challenges of the research community and how they might be met.

The results of the study highlighted the need for transparent and consistent criteria for documentation and metadata in the areas of multilingual and multimodal corpora. The QUEST portal will provide checkers for general metadata standards in common formats such as OLAC¹⁸, COMA¹⁹ or CMDI.²⁰ Specific checkers will also be provided for subject areas where metadata standards are available, such as

¹⁷<https://www.clarin.eu/portal>

¹⁸<http://www.language-archives.org/OLAC/metadata.html>

¹⁹<https://exmaralda.org/en/corpus-manager-en>

²⁰<https://www.clarin.eu/content/component-metadata>

sign language corpora (Crasborn, 2010) and RefCo corpora.²¹ For corpus creators, the Knowledge Base will contain links to common metadata standards and the CLARIN Concept Registry.²²

Another purpose of this study was to identify main points for the development of curation criteria for transcription and annotation. It is clear that for multimodal corpora in general there are no clear criteria, since the field is large and heterogenous. In multilingual corpora there are some conventions which are often used, as described in Section 4.1, and we will provide checkers for these. In addition, some sub-areas have clear annotation conventions, including second-language acquisition research with learner corpora and community interpreting corpora. In these two areas it has been possible to draw up a list of criteria which can be semi-automatically or manually checked. We will also provide links in the Knowledge Base to the CLARIN Resource Families²³ (Fišer et al., 2018), to aid researchers in discovering existing corpora and annotation schemes relevant to their research.

With regards to data storage and protection, the survey participants and experts mentioned two important topics: the difficulty of discovering and following different national and international rules for data protection, and the storage of large amounts of video and audio data, which can be problematic, both in terms of cost and in terms of maintaining long-term storage options. From our survey and expert interviews, it is clear that there is no one rule which must be followed for the anonymisation or pseudonymisation of corpora. The decision depends on the language community involved and the licence under which the data will be made public (see Section 4.2). The QUEST portal will not attempt to check whether any form of anonymisation has been carried out, but will provide links to various resources with information on the ethical, legal, and practical aspects of the decision. The Knowledge Base will contain links to existing resources which can help corpus creators, such as the CLARIN Overview of Data Protection²⁴ and Data Management Plan²⁵ and the DARIAH Consent Form²⁶ wizard.

This mixed study has allowed us to obtain an overall picture of the current needs and challenges in multilingual and multimodal corpus creation and reuse. We have used these findings to inform the development of the QUEST project web portal for quality assurance which aims to improve the reuse potential of such corpora.

Acknowledgements

This work was supported by the BMBF (German Federal Ministry of Education and Research) Project QUEST: Quality-Established. The authors would like to thank the anonymous reviewers for their valuable feedback and Marc Schulder for helpful comments.

References

- Allwood, J. 2008. Multimodal Corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*, volume 1, pages 207–225. Berlin: Mouton de Gruyter. <http://hdl.handle.net/2077/23244>.
- Arestau, E. 2021. Nachhaltige Dokumentation von Metadaten für audiovisuelle Lernerkorpora: Zwischenergebnisse aus dem Projekt QUEST. Poster presented at GAL-Sektionentagung 2021, Würzburg, 15.09. - 17.09.2021.
- Arkhangelskiy, T., Hedeland, H., and Riaposov, A. 2021. Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data. In Navarretta, C. and Eskevich, M., editors, *Selected Papers from the CLARIN Annual Conference 2020*, pages 1–7. Linköping Electronic Conference Proceedings 180. <https://doi.org/10.3384/ecp1801>.
- Arndt, O., Glatz, L., Hummel, B., Porst, M., Schabalowski, W., and Skubatz, S. 2018. Umfrage zum Forschungsdatenmanagement an der FH Potsdam : Projektbericht. Zenodo. <https://doi.org/10.5281/zenodo.1161792>.

²¹<https://doi.org/10.5281/zenodo.6242355>

²²<https://www.clarin.eu/content/clarin-concept-registry>

²³<https://www.clarin.eu/resource-families>

²⁴<https://www.clarin.eu/content/clic-overview-of-data-protection>

²⁵<https://www.clarin-d.net/en/preparation/data-management-plan>

²⁶<https://consent.dariah.eu>

- Bender, E. M. and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604. https://doi.org/10.1162/tacl_a.00041.
- Bergmann, A. 2018. Perspektiven der korpusbasierten Lernaltersprachenanalyse aus fremdsprachendidaktischer Sicht. In Bergmann, A., Caspers, O., and Stadler, W., editors, *Didaktik Der Slawischen Sprachen – Beiträge Zum 1. Arbeitskreis in Berlin*, pages 15–32. Innsbruck University press. <https://doi.org/10.15203/3187-11-5>.
- Crasborn, O. 2010. The sign linguistics corpora network: Towards standards for signed language resources. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA). <https://aclanthology.org/L10-1009>.
- Fandrych, C., Frick, E., Hedeland, H., Iliash, A., Jettka, D., Meißner, C., Schmidt, T., Wallner, F., Weigert, K., and Westpfahl, S. 2016. User, who art thou? User Profiling for Oral Corpus Platforms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 280–287, Portorož, Slovenia, May. European Language Resources Association (ELRA). <https://www.aclweb.org/anthology/L16-1043>.
- Ferus, A., Gorraiz, J., Gründhammer, V., Gumpenberger, C., Maly, N., Mühlegger, J. M., Preza, J. L., Sánchez Solís, B., Schmidt, N., and Steineder, C. 2015. Researchers and their data. Results of an Austria survey—Report 2015. Zenodo. <https://doi.org/10.5281/zenodo.34005>.
- Fišer, D., Lenardič, J., and Erjavec, T. 2018. CLARIN's key resource families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1210>.
- Foster, M. E. and Oberlander, J. 2007. Corpus-Based Generation of Head and Eyebrow Motion for an Embodied Conversational Agent. *Language Resources and Evaluation*, 41(3-4):305–323. <https://doi.org/10/dxk66c>.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., and Crawford, K. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92. <https://doi.org/10.1145/3458723>.
- Hedeland, H., Lehmborg, T., Schmidt, T., and Wörner, K. 2014. Multilingual Corpora at the Hamburg Centre for Language Corpora. In Ruhi, S., Haugh, M., Schmidt, T., and Wörner, K., editors, *Best Practices for Speech Corpora in Linguistic Research*. Cambridge Scholars Publishing, Newcastle upon Tyne. <https://nbn-resolving.org/urn:nbn:de:bsz:mh39-31288>.
- Hedeland, H., Lehmborg, T., Rau, F., Salfner, S., Seyfeddinipur, M., and Witt, A. 2018. Introducing the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). <https://aclanthology.org/L18-1370>.
- Hedeland, H. 2022. FAIR-Prinzipien und Qualitätskriterien für Transkriptionsdaten: Empfehlungen und offene Fragen. In Schwarze, C. and Grawunder, S., editors, *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion Konzepte, Probleme, Lösungen*, pages 346–371. Narr Francke Attempto Verlag GmbH + Co. KG. <https://doi.org/10.24053/9783823394693>.
- Lüdi, G. 2011. Neue Herausforderungen an eine Migrationslinguistik im Zeichen der Globalisierung. *Mobilisierte Kulturen*, 2:15 – 38. <https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-53632>.
- MacWhinney, B. 2000. *The CHILDES Project: Tools for Analyzing Talk*. MacWhinney, B. (2000). 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates. <https://doi.org/10.21415/3mhn-0z89>.
- Rehbein, J., Schmidt, T., Meyer, B., Watzke, F., and Herkenrath, A. 2004. Handbuch für das computergestützte transkribieren nach HIAT. In *Arbeiten Zur Mehrsprachigkeit, Folge b 56*. Universität Hamburg.
- Rubin, H. and Rubin, I. 2005. *Qualitative Interviewing (2nd ed.): The Art of Hearing Data*. SAGE Publications, Inc., Thousand Oaks, California. <https://doi.org/10.4135/9781452226651>.
- Schmidt, T. and Wörner, K., editors. 2012. *Multilingual Corpora and Multilingual Corpus Analysis*, volume 14 of *Hamburg Studies on Multilingualism*. John Benjamins Publishing Company. <https://doi.org/10.1075/hsm.14>.
- Schmidt, T. and Wörner, K. 2014. Exmaralda. In Durand, J., Gut, U., and Kristoffersen, G., editors, *The Oxford Handbook of Corpus Phonology*, pages 402–419. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199571932.013.030>.
- Thomason, S., 2010. *Contact Explanations in Linguistics*, chapter 1, pages 29–47. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781444318159.ch1>.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., and others, . 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018. <https://doi.org/10.1038/sdata.2016.18>.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA). <https://aclanthology.org/L06-1082/>.

Legal Issues Related to the Use of Twitter Data in Language Research

Pawel Kamocki
IDS Mannheim,
Germany
kamocki@ids-
mannheim.de

Vanessa Hanneschläger
OeAW, Austria
vanessa.hanneschlaeger@
oeaw.ac.at

Esther Hoorn
Rijksuniversiteit
Groningen,
the Netherlands
e.hoorn@rug.nl

Aleksei Kelli
University of Tartu,
Estonia
aleksei.kelli@ut.ee

Marc Kupietz
IDS Mannheim,
Germany
kupietz@ids-mannheim.de

Krister Lindén
University of Helsinki,
Finland
krister.linden@
helsinki.fi

Andrius Puksas
Mykolas Romeris University,
Lithuania
andrius_puksas@mruni.eu

Abstract

Twitter data is used in a wide variety of research disciplines in Social Sciences and Humanities. Although most Twitter data is publicly available, its re-use and sharing raise many legal questions related to intellectual property and personal data protection. Moreover, the use of Twitter and its content is subject to the Terms of Service, which also regulate re-use and sharing. The first part of this paper provides an analysis of these issues, whereas the second part discusses two possible strategies to address them: using the new Academic Research product track, which enables authorized researchers to access Twitter API on a preferential basis, or relying on the new statutory copyright exception for Text and Data Mining for research purposes.

1 Introduction

Social media data is useful for a wide variety of research disciplines in Social Sciences and Humanities, such as sociology, computer science, media and communication, political science, and engineering, to name a few. With nearly 400 million users worldwide¹ and over 500 million tweets per day², Twitter is one of the most popular platforms for academic research on social media data.

The main research methods used on social media and Twitter data are: 1) Content Analysis for systematically labelling text, audio, and visual communication from social media; 2) Thematic Analysis

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

¹ After: <https://backlinko.com/twitter-users> (last visit: 4 April 2022).

² After: <https://www.internetlivestats.com/twitter-statistics/#trend> (the number of 500 million tweets per day was already reached in 2013; the same source indicates that as of 4 April 2022 there are on average 9945 tweets per second, which would suggest a much higher number of daily tweets – almost 860 million).

locating patterns within data through data familiarisation, coding, developing and revising themes; 3) Social Network Analysis to measure and map the relationships between individuals, organisations, and other entities; 4) Machine Learning teaching computers with pre-labelled subsets to code the remainder of the data, 5) Discourse-linguistic analyses of the language and treatment of socio-political and other issues in Twitter and the comparison with other media. 6) Semantic Analysis examining the meaning of and the relationship between occurrences of words, phrases, and clauses and 7) Time Series Analysis for plotting the frequency of items or events in the above across time (for further information, see Ahmed 2019).

Although most Twitter data is publicly available, its re-use and sharing (especially in a way compatible with Open Science requirements) raise many legal questions related to intellectual property and personal data protection. Moreover, the use of Twitter and its content is subject to the detailed Terms of Service, which also regulate data re-use and sharing. This paper provides a brief analysis of the above-mentioned issues.

2 Legal Issues in Twitter Data

2.1 Copyright in Tweets

A text is protected by copyright if it is original, i.e., if it constitutes the author's own intellectual creation (CJEU, case Infopaq, C-5/08). Very short texts, such as slogans or titles are often considered unoriginal, because intellectual creation can hardly manifest itself in a very short format. However, the Court of Justice of the European Union ruled that snippets of 11 consecutive words can potentially qualify for copyright protection (idem). This should not be interpreted as a strict measure of originality, but rather as a guideline; on the one hand, not all 11-gramms are protected by copyright, and on the other: some shorter snippets can also qualify for copyright protection. For instance, according to the opinion of advocate general Szpunar (2018) "'All quiet on the Western Front'", declared what is probably the most well-known military report in the history of literature. Featured in the novel by Erich Maria Remarque bearing the same name, this phrase naturally enjoyed, together with the work as a whole, copyright protection'. Kamocki (2020) argues that only n-grams that are no longer than 3 words can safely be regarded as copyright-free.

The maximum length of a tweet is currently set at 280 characters (increased from 140 in November 2017), which corresponds to about 50-60 words in English. This is well enough to be protectible by copyright. However, it has been shown that in practice very few tweets reach the maximum length, and most of them are in fact considerably shorter: an average tweet in English has been reported to be only 33 characters long, i.e. approximately 6-7 words (Perez, 2017). Nevertheless, even this shorter length does not allow to exclude average tweets from copyright protection.

This does not mean that all tweets are indeed original and protected by copyright. Arguably, in reality and from the quantitative perspective most tweets (like 'Big win!', 'LewanGOALski!!!!!!!!!!!!1111' or 'This is crazy LOL') certainly fail to meet the originality criterion. However, a pack of several thousand tweets is likely to contain at least some copyright-protected material (even if it does not include photographs or other media). Since in many cases it is impossible to determine whether a tweet is or is not protected by copyright, it is prudent to consider them as being under copyright. Therefore, especially in analysing tweets en masse, copyright issues have to be observed.

This conclusion has two important implications: one related to the moral rights of authors, the other to the economic rights. Concerning moral rights, Article 6bis(1) of the Berne Convention provides authors with 'the right to claim authorship of the work (a.k.a. paternity right – added by authors) and to object to any distortion, mutilation or other modification of, or other derogatory action in relation to, the said work, which would be prejudicial to his honor or reputation (a.k.a. integrity right – added by authors)'. Even if one sets the integrity right aside (arguably, using a tweet in research is never prejudicial to the author's 'honor or reputation'), the paternity right still obliges researchers to mention the name (or nickname) of the author of every tweet whenever it is quoted or otherwise shared. When tweets are used in bulk, this may lead to a phenomenon known as 'attribution stacking'.

More importantly, the authors have the exclusive right of reproduction and communication to the public of their works. These two economic rights, harmonised in the EU by the 2001 InfoSoc Directive 2001/29/CE (respectively Article 2 and 3), grant the authors of copyright-protected tweets control over

their re-use; in other words, such tweets can only be copied and shared if the author grants permission to do so, or if a statutory exception applies. Both hypotheses are discussed in Section 3 of this paper.

2.2 Tweets as Personal Data

Having established that tweets are potentially copyright-protected, it is now time to examine if they should also be regarded as personal data. Personal data is defined as ‘any information related to an identified or identifiable natural person’ (Article 4, (1) of the GDPR). As per WP29 Opinion 4/2007 on the concept of personal data, information ‘relates to’ a person if it is *about* that person (p. 9).

Tweets necessarily contain information about the author: at the very least the user ID, but possibly also location data or other identifying content (e.g. information about the author’s opinions, preferences, etc.). Therefore, they should be regarded as personal data (see e.g., Gold, 2020) and their processing needs to follow the GDPR, even despite the fact that Twitter is an American company (as per its Article 3.2, the GDPR applies to foreign companies which offer services to EU citizens).

Probably the most important implication of this is the fact that the processing of tweets (including their copying, storage, analysis, anonymisation or any form of sharing) needs a legal basis in order to comply with one of the main principles of data processing under the GDPR, namely the principle of lawfulness (Article 5.1 (a) and Article 6 of the GDPR).

Contrary to a common misconception, consent of the data subject (i.e., the person that the data refers to) is not always necessary; it is only one of the available options. Moreover, consent does not have to be given in writing or even (in principle) be explicit, it can also be implied, inferred from an unambiguous affirmative action. Since Twitter provides its users with the possibility to fine-tune their privacy settings, including public availability of their tweets and profile information, mere making tweets publicly available may arguably be interpreted as granting consent to their processing for research purposes, taking into account that Twitter also expressly informs the users (in its Rules and Policies) that it conducts research on data. A problem with this approach arises, however, when the user deletes a tweet, or changes its parameters in such a way that it is no longer public. This should probably be interpreted as withdrawal of consent (under the GDPR, consent can be withdrawn at any time, cf. Article 7(3) of the GDPR). In such a case, the processing of such tweets should stop (see: EDPB Guidelines 05/2020 on consent under Regulation 2016/679, para. 117), and they should be deleted from the corpus, which is a major (at least organisational) obstacle.

As mentioned above, alternative legal bases are also available. One of such alternatives, often regarded as compatible with research purposes, is ‘legitimate interest of the controller’ (Article 6.1(f) of the GDPR). In order to rely on this basis, the controller should perform a ‘balance of interests’ test to assess whether the interests of the data subject do not override the controller’s interest (e.g. in using the data for research purposes). According to the WP29 Opinion 06/2014 (p. 55), the balancing test should take into account such elements as the reasonable expectations of the data subject (cf. also Recital 47 of the GDPR), the nature of the data, and the potential impact of the processing on the data subject. In the context of language research on Twitter data, it seems that the outcome of the balancing test will likely be in favour of the processing, considering that the data in question are short messages made public by the data subject, that the data subject should be aware that they can be used for research purposes, and that language research is highly unlikely to have any negative impact on the data subject. However, also in this case it can be argued that in a situation where the data subject subsequently deletes the tweet or restricts access to it (making it invisible to the general public), the balance is tilted in the opposite direction, and the controller can no longer rely on the ‘legitimate interest’ legal basis to process the tweet. Furthermore, when the processing is based on ‘legitimate interest’, the data subject has the right to object to the processing (Article 21 of the GDPR) – in such a case, the processing can continue only if it passes a stricter test for ‘compelling legitimate grounds’. For now, in the absence of any EDPB guidelines on the right to object, still relatively little is known about this right and the consequences of its exercise.

Yet another legal basis that can be relevant for the processing of Twitter data for language research purposes is ‘public interest’ (Article 6.1(e) of the GDPR). In order to be able to rely on this ground, processing has to be based on an interest clearly laid down in the law (typically, this basis is used e.g. by tax authorities). In some CLARIN countries, such as Finland or Norway, whose national laws contain specific provisions to this effect, this basis is available and recommended for researchers.

Rather exceptionally, tweets may also contain special categories of personal data (the so-called ‘sensitive data’, cf. Article 9 of the GDPR), i.e. data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, as well as genetic data, biometric data, and data concerning health, sex life or sexual orientation. The processing of such data must be based on specific, strictly defined legal bases (for example, consent for the processing of such data has to be ‘explicit’, and so our analysis of ‘consent’ as a legal basis for processing Twitter data for research purposes cannot apply). However, one of these specific legal bases for the processing of sensitive data applies to Twitter data: according to Article 9.2(e), sensitive data can be lawfully processed if they have been ‘manifestly made public by the data subject’. Rather obviously, publicly tweeting about e.g., one’s health (by announcing an operation or a diagnosis) does count as making this information ‘manifestly public’, and therefore the above-mentioned legal basis can apply. However, it is much less obvious whether it continues to apply after the relevant tweet has been deleted or restricted. In any case, given the sensitive nature of those special categories of personal data, researchers should always be very prudent while relying on this legal basis.

Tweets may also contain personal data related to third persons, i.e. individuals other than the author of the tweet. Although a lot depends on the circumstances of every specific case, in general it is rather difficult to find an appropriate legal basis for the processing of such data. In particular, it seems difficult to rely on consent (as usually nothing indicates that the third person has consented to her personal data being published in the tweet), or legitimate interest (as the third person has little reason to expect that her data, tweeted by someone else, will be used for research). Unfortunately, it is rather impossible to automatically detect tweets containing third persons’ personal data, which further complicates the use of Twitter data for language research purposes.

Even if the processing complies with the principle of lawfulness (i.e., it has an appropriate legal basis), there is a number of other requirements in the GDPR that it has to meet. One of such requirements is related to the principle of transparency, under which the data subjects should be provided with information about the processing in a concise, transparent, intelligible and easily accessible form (cf. Articles 12 and 14 of the GDPR). This information includes, inter alia, the identity and contact details of the data controller, the purposes for which the data are being processed, the data retention period, and the rights of the data subject (for more information, see WP29 Guidelines on transparency (WP260rev01)). In the context of Twitter data analysis, taking into account the sheer amount of processed data and concerned data subjects, this principle seems particularly difficult to observe. However, the GDPR, in its Article 14.5(b), includes an exception from this principle for cases where provision of the information proves impossible or would involve disproportionate effort. As per the Article 14.5(b) itself, this exception can apply in particular to the processing carried out for research purposes. In assessing whether the necessary effort is disproportionate, according to the Recital 62 of the GDPR, account should be taken of such elements as the number of data subjects (the higher the number, the bigger the effort), the age of the data (the older the data, the bigger the effort) and any appropriate safeguards adopted by the controller (e.g. pseudonymisation, encryption, restricted access to the collected data, etc.). This assessment has to be made on a case-by-case basis, but it seems that when analysing Twitter data for language research purposes at least the first element – the number of data subjects – will generally weigh in favour of the exception of ‘disproportionate effort’. It should be noted that even if the exception applies, the controller should still ‘take appropriate measures to protect the data subject’s rights and freedoms and legitimate interests, including making the information publicly available’. A reasonable solution would be to publish the relevant information e.g. on the project’s website.

2.3 Tweets and Contracts

In order to be able to tweet, one needs to create a Twitter account and accept (among other documents, such as the Privacy Policy) Twitter’s Terms of Service (ToS)³. Upon acceptance, the ToS become a binding contract that both the user and the platform provider are bound to respect.

It has been demonstrated *supra* that tweets can be protected by copyright. Therefore, it is particularly interesting for further analysis to examine how copyright issues are addressed in the ToS. In Section 3

³ Available at <https://twitter.com/en/tos#intlTerms> (last visit: 9 February 2022).

of the ToS, the paragraph entitled ‘Your Rights and Grant of Rights in the Content’ provides the following:

‘By submitting, posting or displaying Content on or through the Services, [the user] grant[s] [Twitter] a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods now known or later developed (for clarity, these rights include, for example, curating, transforming, and translating). This license authorizes [Twitter] to make [the user’s] Content available to the rest of the world and to let others do the same’.

This means that although the user retains copyright in his tweets, he grants Twitter a very broad permission (license) to re-use them for free on a non-exclusive basis (i.e., the user can still use the tweets himself and authorise others to do so). As a consequence, someone who would like to copy and share tweets can receive the necessary authorisation either directly from the user (which in most cases is unworkable in practice, given the sheer number of Twitter users) or from Twitter (a license from Twitter would be a *sublicense*; sublicensing is explicitly authorised by the ToS). Also, theoretically, nothing prevents the users from re-publishing their own tweets outside of Twitter, including e.g. in .xml format, and under an open license.

Twitter ToS also grants every user access to the Twitter Services, but no general sub-license to re-use the content (a limited personal license is provided only to use the software provided as part of the Services, for the sole purpose of enabling the user to enjoy the services). Moreover, certain uses and actions are expressly forbidden: for example, the user is not allowed to ‘access or search or attempt to access or search the Services by any means (automated or otherwise) other than through (...) currently available, published interfaces that are provided by Twitter (and only pursuant to the applicable terms and conditions)’. Interestingly, it is allowed to crawl the Services (i.e., presumably, to use Twitter to find, discover and visit URLs) in accordance with the provisions of the robot.txt file (such uses can hardly be efficiently prohibited); it is not allowed, however, to scrape the Services (i.e. to extract the data) without Twitter’s prior permission. It seems therefore that mining of tweets without specific permission, even if done for research purposes, would violate Twitter ToS, which may lead to suspension or termination of the user account(s) that is (are) at the origin of these actions, or perhaps even to a lifetime ban. Theoretically, Twitter could also sue for damages for breach of contract, but this, in our opinion, is highly unlikely to happen, for at least two reasons 1) the economic loss suffered by Twitter would probably be negligible, if even possible to quantify, and therefore the amount of compensatory damages that could be claimed by Twitter would also be negligible; 2) suing a non-commercial research organisation such as a university would result in controversies that may be harmful to the image of the company. Despite its low probability, the fear of legal action from Twitter is probably the reason why those researchers who have indeed scraped data from Twitter are not transparent about it, e.g. in their ethical self-assessments, and therefore many Twitter corpora remain underexplored and ‘under the radar’.

Although the authors have not tested this in practice, they assume that scraping tweets is not only forbidden by Twitter ToS, but also made impossible (or at least very difficult) by technological protection measures (TPM). For example, Twitter may detect ‘unhuman’ use of its Services (such as consulting a very large amount of URLs in a very short period of time from one IP address) and prevent it. Circumventing such technological measures is not only expressly prohibited by the Twitter ToS, but also in principle forbidden by law (cf. Article 6 of the InfoSoc Directive).

Another provision of the ToS that is noteworthy from the point of view of this analysis is related to the termination of the contract. In Section 4, the paragraph entitled ‘Ending these Terms’ provides that Twitter may suspend or terminate the user’s account or cease providing the user with all or part of the Services at any time ‘for any or no reason’ [sic!]. Therefore, there is no legal guarantee that Twitter will continue to provide its Services in the future, which is a major problem from the point of view of sustainability of Twitter data used for research purposes.

3 Using Twitter Data for Language Research – Possible Strategies

In the previous section, it has been demonstrated that in order to be able to lawfully scrape and analyse tweets for language research purposes, researchers need, in addition to observing GDPR principles, to

obtain a specific permission from Twitter, or alternatively to rely on a statutory exception. Both solutions have their advantages and disadvantages, as discussed in this section of the paper.

3.1 Twitter API for Academic Research

The use of Twitter data for language research purposes is still associated with considerable organisational effort and lack of legal certainty. In this context, simply applying for a specific permission from Twitter may be a reasonable solution. This could not only clear any copyright-related issues, but also diminish the burden related to the GDPR -- when the processing is carried out solely through an API provided by Twitter, it can be argued that Twitter is a joint controller for the processing. In July 2020, Twitter launched a new version (v2) of its API. Reportedly, academic researchers were one of the largest groups of the API users; for this reason, in January 2021 Twitter has launched a new Academic Research product track, allowing for a preferential access to the API (Tornes and Trujillo, 2021).

In theory, the Academic Research track allows for a 10 000 000 monthly tweet volume cap (compared to 500 000 in the general track), although this also depends on the streaming endpoint limits which reportedly are not entirely up to this standard yet (although they are expected to be raised soon). Moreover, it is also possible to use more detailed queries and rules (1024 characters per query/rule in the Academic Research product track, as opposed to 512 in the Standard track). In addition, the Twitter Development Agreement allows academic researchers to distribute an unlimited number of Tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research (otherwise, 'only' 1 500 000 Tweet IDs per 30 day period can be shared). The content itself, however, cannot be shared. This might be seen as an inconvenience, but it allows to solve many GDPR-related problems with tweets that have been deleted or restricted by the user (see above).

The Academic Research product track is available to: 1) researchers, post-docs, professors of fellows at academic institutions (undergraduates are expressly excluded); 2) Master's students working on theses; 3) PhD candidates working on dissertations and 4) persons affiliated with an academic institutions and working on a clearly defined research project. In all cases, the applicant has to pursue a non-commercial purpose, and have a Twitter account.

In the process of applying for the Academic Research track, the applicant has to prove his or her affiliation with an academic institution (by providing a link to the webpage on his or her institution's website listing his or her name, or to his or her Google Scholar profile), provide information about the institution, his or her department or lab, and his or her current role in the research group. Then, the applicant is asked to answer a very detailed questionnaire about his project including questions about its name, funding, methodology, the planned use of Twitter data and ways of sharing the outcomes. Arguably, some may see this questionnaire as intrusive and unacceptable from the point of view of academic freedom.

Access to the Track is free. There is no information available as to how many requests are granted, and what are the admission criteria. Successful candidates are bound, like anyone with access to the API, by the Twitter Development Agreement and Policy. These documents strictly prohibit any attempt to exceed or circumvent access limitations (rate limits). Moreover, Twitter retains the right to immediately terminate or suspend access to the API at any time and for any reason. It can be expected that any attempt to exceed the permissions granted by Twitter will be met with termination of access to the API. This, combined with the possibility for Twitter to modify or stop providing its Service at any time, is far from optimal from the point of view of sustainability of research data accessed via the API.

3.2 Statutory Exception for Text and Data Mining for Scientific Research Purposes

As explained above, re-use of copyright-protected content is only possible if it is authorised by the author (directly or indirectly), or if it is exempted from authorisation by a statutory exception. Recently the Directive 2019/790 on copyright in the Digital Single Market (DSM Directive) introduced (in its Article 3) a new statutory exception, supposed to cover such scenarios as using Twitter data for language research purposes at research institutions. The Directive is now transposed in most EU Member States (the deadline for transposition was set for 7 June 2021, but in many countries the relevant legislative processes were delayed by the COVID-19 pandemic).

This new exception allows research organisations (such as universities) and cultural heritage institutions (such as libraries, museums or archives) to make copies of copyright-protected content in

order to carry out text and data mining for scientific research purposes (including in public-private partnerships). The exception only applies to content to which the above-mentioned institutions have ‘lawful access’. This requirement is often presented as a hurdle, but when it comes to publicly available tweets, the criterion is easily met: as per Recital 14 of the DSM Directive, ‘Lawful access should also cover access to content that is freely available online’.

In general, copyright exceptions are overridden by contracts. In other words, if a contract (such as Terms of Service) prohibits certain uses (like scraping), this prohibition remains in principle unaffected by statutory exceptions. However, the exception for text and data mining for research purposes has a rare and very important feature: it is not overridable by contracts, i.e. any contractual provision contrary to this exception is unenforceable (Article 7.1 of the DSM Directive). This means that the beneficiaries of the exception (research organisations and cultural heritage institutions) may scrape Twitter data, despite the general prohibition of scraping in the Twitter ToS. It remains to be seen if such use will be tolerated by Twitter which, as per the ToS, can cease to provide the Services to any user for any reason, including for no reason at all (see above). In this context, the exception can shield against a legal action from Twitter for breach of contract, but not against unilateral termination of the contract by Twitter.

Another aspect of the exception concerns its relation with technological protection measures. This seems to be the biggest grey area of the exception, as Article 3.3 of the DSM Directive allows platform providers to apply technological measures to disable text and data mining for research purposes, but only to the extent necessary to ensure the security and integrity of their networks and databases. In our opinion, Twitter would have a good chance to succeed in arguing that TPMs implemented to prevent unauthorised scraping are, in fact, necessary to achieve such goals, as unlimited scraping might place too heavy a burden on their servers and affect the accessibility of their services for other users; however, it still remains to be seen how this issue will be worked out in practice. According to the DSM Directive, Member States shall encourage stakeholders to define commonly agreed best practices in this area.

The copies made under the exception (i.e., corpora) have to be stored ‘with the appropriate level of security’ to protect them against unauthorised access. They can, however, be re-used in other projects or for evaluation purposes. Unfortunately, the exception itself does not seem to allow any sharing of the data, although there might be slight variations between implementations in the various EU Member States; for example, the German implementation (Section 60d of the German Copyright Act) allows for the corpus to be shared with a limited circle of persons for joint scientific research, which seems to allow sharing within research infrastructures such as CLARIN.

The advantage of relying on this exception rather than the Twitter API in using Twitter data for language research purposes is a greater degree of autonomy and control over the data collection process. On the other hand, it remains an uncharted territory with many great areas. Unlike the use of Twitter APIs, the statutory exception also does not provide any relief regarding the GDPR compliance.

4 Conclusion

Twitter data present a number of legal issues: tweets can be protected by copyright, they contain personal data, and access to them is regulated by the Terms of Service. This, however, does not mean that Twitter data are out of reach for language researchers. Quite the contrary, there are at least two ways to get hold of such data: via the API provided by Twitter (with a specific, preferential track dedicated to academic research), or by relying on the new copyright exception for Text and Data Mining. None of these approaches is fully satisfactory. Moreover, taking into account the specificities of national laws (especially with regards to the Text and Data Mining exception), identified research questions and adopted research methods, specific solutions for handling Twitter data should still be adopted on a case-by-case basis.

One such solution that looks quite tempting might be a hybrid approach between the Twitter API and the statutory exception; in this scenario, the data are accessed via the Twitter API, then copied on the basis of the copyright exception for Text and Data Mining, anonymised, stored, re-used and potentially even shared with other researchers (this last element seem to depend mostly on the applicable national law). As both the Academic Research track in the Twitter API and the relevant copyright exception were only introduced relatively recently, best practices have yet to emerge – also by trial and, quite inevitably, by error.

References

- Ahmed, W. 2019. Using Twitter as a data source: an overview of social media research tools (2019). Available at <https://blogs.lse.ac.uk/impactofsocialsciences/2019/06/18/using-twitter-as-a-data-source-an-overview-of-social-media-research-tools-2019/> (09.02.2022).
- Gold, N. 2020. *Using Twitter Data in Research. Guidance for Researchers and Ethics Reviewers*. University College London. Available at <https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf> (09.02.2022).
- Kamocki, P. 2020. When Size Matters. Legal Perspective(s) on N-grams. Proceedings of CLARIN Annual Conference 2020. 05 – 07 October 2020. Virtual Edition. Ed. Costanza Navarretta, Maria Eskevich. CLARIN, 166-169. Available at https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf (09.02.2022).
- Perez, S. 2017. Twitter officially expands its character count to 280 starting today. Available at <https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/?guccounter=1> (09.02.2022).
- Szpunar, M. 2018. Opinion of Advocate General Szpunar delivered on 25 October 2018. Case C-469/17. Funke Medien NRW GmbH v Bundesrepublik Deutschland. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1574794094419&uri=CELEX:62017CC0469#t-ECR_62017CC0469_EN_01-E0002 (09.02.2022).
- Tornes, A., Trujillo, L. 2021. Enabling the future of academic research with the Twitter API. Available at https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.html (09.02.2022).

The Interaction of Personal Data, Intellectual Property and Freedom of Expression in the Context of Language Research

Aleksei Kelli
University of Tartu,
Estonia
aleksei.kelli@ut.ee

Krister Lindén
University of Helsinki,
Finland
krister.linden@helsinki.fi

Pawel Kamocki
IDS Mannheim,
Germany
kamocki@ids-mannheim.de

Kadri Vider
University of Tartu,
Estonia
kadri.vider@ut.ee

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Ramūnas Birštonas
Vilnius University,
Lithuania
ramunas.birstonas@tf.vu.lt

Vadim Mantrov
University of Latvia,
Latvia
vadims.mantrovs@lu.lv

Vanessa Hanneschläger
OeAW, Austria
vanessa.hanneschlaeger@gmail.com

Riccardo Del Gratta
ILC, Italy
riccardo.delgratta@ilc.cnr.it

Age Värv
University of Tartu,
Estonia
age.varv@ut.ee

Gaabriel Tavits
University of Tartu,
Estonia
gaabriel.tavits@ut.ee

Andres Vutt
University of Tartu,
Estonia
andres.vutt@ut.ee

Esther Hoorn
University of Groningen,
The Netherlands
e.hoorn@rug.nl

Jan Hajic
Charles University,
Czechia
hajic@ufal.mff.cuni.cz

Arvi Tavast
Institute of the
Estonian Language,
Estonia
arvi@tavast.ee

Abstract

Language researchers are usually aware of intellectual property and personal data (PD) requirements. The problem, however, arises when these two legal regimes have conflicting requirements. For instance, when copyright law requires the acknowledgement of the author, but personal data law enshrines the data minimisation principle. It is a practical question for a language researcher whether he should name the author of the text used for, e.g., building a language model, or follow the data minimisation principle not to name the author.

The access right that a data subject has introduces similar conflicts. The question is what the scope of the access right is. Does it cover only processed personal data, or does it extend to data derived from PD?

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

Aleksei Kelli, Krister Lindén, Pawel Kamocki, Kadri Vider, Penny Labropoulou, Ramūnas Birštonas, Vadim Mantrov, Vanessa Hanneschläger, Riccardo Del Gratta, Age Värv, Gaabriel Tavits, Andres Vutt, Esther Hoorn, Jan Hajic Charles and Arvi Tavast 2022. The Interaction of Personal Data, Intellectual Property and Freedom of Expression in the Context of Language Research. *Selected papers from the CLARIN Annual Conference 2021*. Ed. by Monica Monachini and Maria Eskevich. Linköping Electronic Conference Proceedings 189, pp. 76–87. DOI: <https://doi.org/10.3384/9789179294441>

The interaction of the freedom of expression with PD protection entails several problems. The question is whether researchers can publish their research results containing personal data. The General Data Protection Regulation establishes a general framework that needs to be implemented by EU member states. We analyse different implementations based on examples from several EU countries.

1 Introduction

There is an awareness that intellectual property¹ and personal data² (PD) protection are relevant in language research. These two regimes are often applicable simultaneously, and their requirements might seem contradictory. Therefore, we have chosen three specific cases³ to outline the interaction of intellectual property and personal data protection and provide preliminary guidance.

Firstly, we explore the interplay between the data minimisation principle and the right to be acknowledged as the author (the attribution/paternity right). On the one hand, the data minimisation principle enshrined in the General Data Protection Regulation (GDPR) requires processing⁴ as little personal data as possible (Art. 5 (1) c)). According to the European Data Protection Board (EDPB) “Data minimisation substantiates and operationalises the principle of necessity” (2019: 21). On the other hand, the Berne Convention Art. 6^{bis}, which sets the international standard and binds all current EU member states (and almost all of the remaining world), gives authors the attribution (paternity) right. The relevant question here is whether a researcher who has collected language data containing copyrighted content (for further discussion on the process of development of language technologies from the legal perspective, see Kelli et al 2020) should attribute the author of the content or follow the data minimisation principle and remove all personal data (e.g., the author’s name) that is not necessary for processing.

The second case concerns intellectual property protection and the data subject’s access right. A researcher might need to decide what data the access right covers in practical terms. Is it only raw personal data⁵ or personal data derived from raw personal data?

Thirdly, we discuss the impact of personal data protection on freedom of expression since publications constitute research outcomes. The two previous questions do not require comparative analysis, but the situation is different in this case. Therefore, we rely on the General Data Protection Regulation implementation model of the following European Union countries: Austria, Czechia, Estonia, Finland, France, Germany, Greece, Italy, Latvia, Lithuania, and the Netherlands. Even though not all EU countries are studied, we can draw preliminary conclusions about the implementations. Adding more countries would not change the general picture.

¹ Intellectual property can be defined as “rights resulting from intellectual activity in the industrial, scientific, literary or artistic fields”. Art. 2 of the Convention Establishing WIPO. IP is traditionally divided into three main categories: 1) copyright; 2) related rights to copyright; 3) industrial property.

² The General Data Protection Regulation (GDPR) defines personal data as “any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person” (Art. 4 (1)).

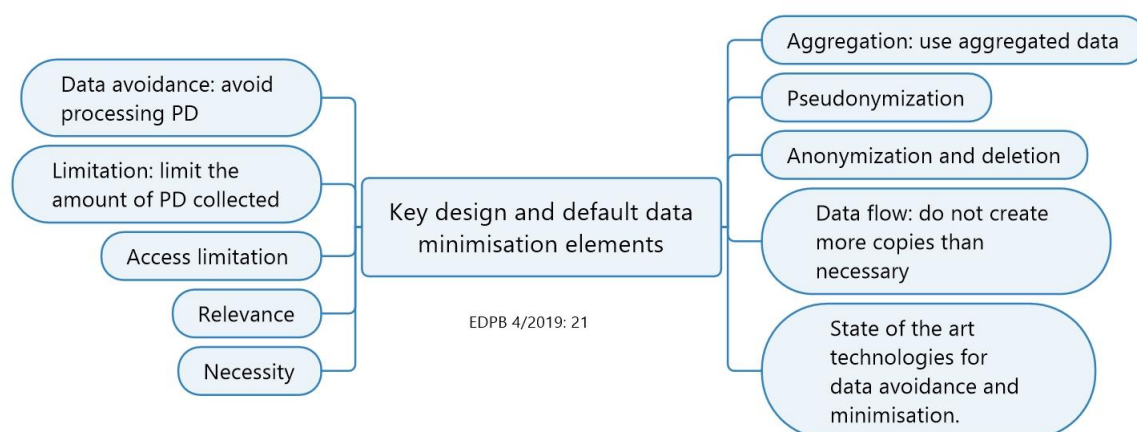
³ It should be mentioned that there are several IP and PD protection interaction points whose systematic mapping is outside the scope of this article. Therefore, we chose cases that could potentially be relevant for language researchers.

⁴ The GDPR defines processing of personal data as “any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction” (Art. 4 (2)).

⁵ In the context of this article, the concept of raw personal data refers to information such as age, height, weight, nationality, income, physical characteristics and so forth. Derived personal data is based on raw personal data (e.g., subject’s profile as a consumer, the estimation of person’s life expectancy and so forth).

2 The data minimisation principle and the right of attribution

According to the data minimisation principle, PD must be “adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed” (GDPR Art. 5 (1) clause c). The European Data Protection Board (EDPB) outlines the data minimisation obligation for default processing with reference to GDPR Art. 25 (2) as containing the following elements: 1) amount of personal data collected (unnecessary data is not collected); 2) the extent of their processing (processing is limited to what is necessary); 3) the period of their storage (the retention period is no longer than necessary); 4) their accessibility (the access is limited to what is necessary) (2019: 12-14). The following graph visualises the data minimisation principle as conceptualised by EDPB (2019: 21):



The focus of the article is not on the data minimisation principle as such but on the identification of the data subject. The European Data Protection Board is of the following opinion: “Minimising can also refer to the degree of identification. If the purpose of the processing does not require the final set of data to refer to an identified or identifiable individual (such as in statistics), but the initial processing does (e.g. before data aggregation), then the controller shall delete or anonymise personal data as soon as identification is no longer needed. Or, if continued identification is needed for other processing activities, personal data should be pseudonymised to mitigate risks for the data subjects’ rights” (2019: 21).

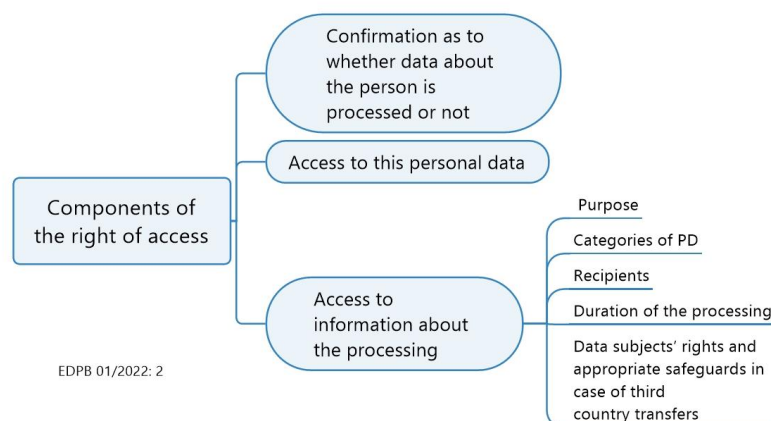
Pursuant to Art. 6^{bis} (1) of the Berne Convention, “the author shall have the right to claim authorship of the work”. The InfoSoc Directive also contains the obligation to identify the source (incl. the author’s name), e.g., in the context of quotation or research exceptions (Art. 5 (3), esp. (a) and (d)). The EU case law reiterates the obligation (e.g., C-145/10). Copyright laws of different European Union member states contain the same principle. For instance, according to the Estonian Copyright Act “The author of a work has the right to appear in public as the creator of the work and claim recognition of the fact of creation of the work by way of relating the authorship of the work to the author’s person and name upon any use of the work (right of authorship)” (§ 12 (1) clause 1). In other words, there is a legal obligation to acknowledge the author of a work. Therefore, it is compatible with the GDPR since it names compliance with a legal obligation as a legal basis for PD processing (Art. 6 (1) c)).

An overarching theme for this and the following section concerns legal obligations relating to derived data, e.g., data derived through text and data mining (TDM). For further discussion, see Kelli et al. (2020). Interestingly, the TDM exception contained in the DSM Directive does not require attribution. However, it should be borne in mind that the TDM exception only limits the reproduction right and does not allow any communication to the public. If the results of TDM are disseminated (e.g., based on quotation or research exception), the attribution right has to be honoured.⁶

⁶ The attribution right exists only in case of the existence of copyrighted content. In the EU case law, it is pointed out that 11 consecutive words could be copyright protected (C-5/08). However, it does not say that less than 11 words are not copyrighted. For further discussion, see Kamocki 2020.

3 The right of access and intellectual property protection

In combination with the right to be informed and the principle of transparency, the access right forms a foundation for exercising the data subjects' rights. The access right requires the controller to provide information on the processing of PD, as well as access to the data (GDPR Art. 15). The European Data Protection Board conceptualises the right of access as follows (2022: 2):



The first question for research organisations and researchers (data controllers) is the scope of the access right. The question is whether the access right applies to raw personal data or personal data derived from raw personal data. In other words, this question asks what PD covers.⁷ The Court of Justice of the European Union (CJEU) has not been remarkably consistent. For instance, it has explained that “There is no doubt that the data relating to the applicant for a residence permit and contained in a minute, such as the applicant’s name, date of birth, nationality, gender, ethnicity, religion and language, are [...] ‘personal data’ [...] As regards, on the other hand, the legal analysis in a minute, it must be stated that, although it may contain personal data, it does not in itself constitute such data” (C-141/12 paragraphs 38, 39). In another case, the CJEU held that “the written answers submitted by a candidate at a professional examination and any comments made by an examiner with respect to those answers constitute personal data” (C-434/16).

Understandably, the concept of personal data should be interpreted consistently and extensively. However, there is no legal clarity on whether data derived from PD should be made available. WP29 (2016: 9) suggests in the context of the right of portability (see GDPR Art. 20) that “user categorisation or profiling are data which are derived or inferred from the personal data provided by the data subject, and are not covered by the right to data portability”.

However, in its very recent draft guidelines, the EDPB (2022, paragraph 96) clearly states that not only the raw data provided by the data subject but also personal data derived and inferred from such data should be provided to the data subject who requests access to his or her personal data. It should be noted that non-personal data, even derived or inferred from the data subject’s personal data, are not concerned with such requests. The data surrounding PD does not have to be made available as well.

Within the context of language research, the question is whether the data subject could require access to a language model trained using his PD. First, a model that does not contain any personal data is not concerned with the right of access. Moreover, the language model containing PD can be protected by intellectual property rights (database copyright, database *sui generis* right and trade secret⁸). The rightholder should have an exclusive right to decide who can access it. The GDPR accommodates this line of argument in its Recital 63, explaining the nature of the access right: “That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in

⁷ For the concept of PD, see WP29 2007.

⁸ Article 21 of the trade secrets directive defines a trade secret as information not generally known, having commercial value and its holder has taken steps to keep it secret.

particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject”.

To sum up, the access right does not cover access to a language model containing PD as a whole, especially when this model can be considered a trade secret. In this context, IP rights prevail over data protection.

4 Data subject’s rights and freedom of expression

4.1 General background

The data subject has the right to object to the processing and obtain the erasure, restriction or rectification of PD concerning himself (GDPR Art. 16, 17, 18, 21). These rights may conflict with the author’s right to make his work available. This question can be framed as an interaction of personal data protection and freedom of expression (FoE). Personal data protection is not an absolute right (GDPR Rec. 4). Furthermore, freedom of expression is guaranteed by all major international human rights treaties and European legal acts, such as the Universal Declaration of Human Rights (Article 19), the Convention for the Protection of Human Rights and Fundamental Freedoms (Article 10), the Charter of Fundamental Rights of the European Union (Article 11). Therefore, the GDPR allows Member States to limit the data subject’s rights to reconcile PD protection with the freedom of expression. According to GDPR Rec. 153, “This should apply in particular to the processing of personal data in the audiovisual field and in news archives and press libraries”.

There are two intriguing questions concerning the interaction of personal data protection and freedom of expression:

1) how to strike a fair balance between personal data protection and freedom of expression in research settings? Freedom of expression is usually framed in the context of newspapers publishing facts about public figures. Freedom of academic expression is somewhat unclear. Still, it has been interpreted to apply to, e.g., x-ray pictures of medical case studies as standard practice. Such accompanying material is publicly disclosed in a scientific journal to illustrate the published case. There is no need to obtain the consent of the x-rayed person for this purpose. Usually, a person cannot be directly identified from such an x-ray. However, if the medical condition is rare, the individual may still be identifiable with the help of additional information. As the GDPR defines data concerning health as special categories of PD (Art. 9), this is an especially delicate example.

2) how and where to draw a line between processing for academic expression and research purposes. Research publication requires prior research. The question is whether this research is covered with the freedom of academic expression. We admit that the processing could be covered by the freedom of expression except when data is present in the research publication. It is important to emphasise that the principles of data minimisation, purpose limitation, accuracy, fairness (GDPR Art. 5), and other requirements need to be followed. Research quality and funding conditions often require the publication of research data to ensure reproducibility and verifiability of research results. Therefore, there is tension between the requirements on providing open data and protecting personal data. For further discussion, see Kelli et al. (2018).

4.2 Implementation models of EU countries to strike a fair balance between freedom of expression and personal data protection

The main aim of the General Data Protection regulation is to create a uniform regulatory framework throughout the European Union. However, some aspects of personal data protection are delegated to the EU member states. Striking a fair balance between freedom of expression and processing personal data is one of them.

According to Art. 85 (1) of the GDPR “Member States shall by law reconcile the right to the protection of personal data pursuant to this Regulation with the right to freedom of expression and information, including processing for journalistic purposes and the purposes of academic, artistic or literary expression”. The GDPR Art. 85 (1) further specifies: “For processing carried out for journalistic purposes or the purpose of academic artistic or literary expression, Member States shall provide for exemptions or derogations from Chapter II (principles), Chapter III (rights of the data subject), Chapter IV (controller and processor), Chapter V (transfer of personal data to third countries or international organisations), Chapter VI (independent supervisory authorities), Chapter VII (cooperation and consistency) and Chapter IX (specific data processing situations) if they are necessary to reconcile the right to the protection of personal data with the freedom of expression and information”.

Analyzing the implementation routes of selected EU countries to guarantee academic freedom of speech exemplifies several differences. The approach of the EU countries varies from countries that do not have any specific provisions to countries with a very detailed regulatory framework. Some countries are placed between the two.

4.2.1 Countries without specific provisions

For instance, the **German** Federal Data Protection Act (BDSG) does not contain any rules specifically implementing Article 85 of the GDPR. The existing broad derogation for research and archiving purposes (Article 27 of the BDSG) is based on Article 89, not 85 of the GDPR. It seems to be deemed sufficient by the legislator (Deutscher Bundestag, 2018). Specific state acts regarding media and journalistic expression exist in many federal states (Länder), e.g., Hessisches Pressegesetz or Landesmediengesetz Baden-Württemberg.

4.2.2 Countries with a general provision

Several studied countries have a general provision (with minor additions) limiting the applicability of the General Data Protection Regulation to protect freedom of expression. These countries are **Austria** (the Austrian Data Protection Amendment Act), **Finland** (the Finnish Data Protection Act), **Latvia** (the Latvian PDPA), and **Lithuania** (the Republic of Lithuania Law on Legal Protection of Personal Data). These general provisions as such are not very informative and, due to their similar character, are not presented here.

Although these countries have a literal implementation model, some still have some additional norms. For instance, Art. 7(2) of the Republic of **Lithuania** Law on Legal Protection of Personal Data further provides that the Inspector of Journalist Ethics shall monitor the application of the GDPR and the Law on Legal Protection of Personal Data and ensure that this legislation applies to the processing of personal data for journalistic purposes and academic, artistic or literary purposes. Therefore, deviating from the general principle, the Inspector of Journalist Ethics, not the State Data Protection Inspectorate, is responsible for supervising the processing of personal data for journalistic purposes and academic, artistic or literary purposes. Currently, the Lithuanian case law and the public decisions made by the Inspector of Journalist Ethics is too fragmented to make any conclusive statements about the interplay of the FoE and PD protection, but it seems that the intention is to interpret the concepts of “journalistic purposes” and “academic, artistic or literary purposes” broadly, as foreseen in Recital 153 of the GDPR.

Article 32(3) of the **Latvian** PDPA states that when processing data for academic, artistic or literary expression, provisions of the GDPR (except for Article 5) shall not be applied if all of the following conditions are present: 1) Data processing is conducted by respecting the right of a person to private

life, and it does not affect interests of a data subject which require protection and override the public interest; 2) Compliance with the provisions of the GDPR is incompatible with or prevents the exercise of the rights to freedom of expression and information.

As one may observe from this quoted provision, it is essentially based on Article 85 (2) of the GDPR and contains two parts. The Latvian legislator reacted to a necessity to provide exemptions or derogations from certain chapters of the GDPR. As one may observe from the phrase ‘except for Article 5’ contained in the quoted provision, the Latvian legislator chose to apply Article 5 (i.e. principles relating to the processing of personal data) to be observed while processing data for academic, artistic or literary expression from all provisions included in the relevant chapters of the GDPR. At the same time, the Latvian legislator provided two cumulative preconditions referred to in the quoted provision for processing data for academic, artistic or literary expression in order to avoid the application of the rules included in relevant chapters of the GDPR. Therefore, if at least one of these two cumulative preconditions is not met, the GDPR in full should be applied for such processing.

4.2.3 Countries with an elaborate provision

There are also countries with a more elaborate approach to the interaction of personal data protection and freedom of expression, such as **Czechia, Estonia, France, Greece, Italy** and the **Netherlands**.

Czechia seems to be the EU country with the most detailed regulation. In **Czechia**, the GDPR is implemented by a completely new law (No. 110/2019 Coll.). Art. 17 gives the legal basis for personal data processing for journalistic, academic, artistic or literary expression: (1) Personal data may also be processed if it serves, in a reasonable manner, journalistic purposes or purposes of academic, artistic or literary expression. (2) The processing of personal data for the purposes referred above is not subject to authorisation or approval of the Office and enjoys the right to protection of the source and content of information, even in the case of the processing of personal data in a manner allowing remote access.

Articles 18 to 22 are devoted to the exceptions related to the right of the subject to be informed, exceptions related to the protection of the source and contents of the personal information, exceptions to the right for corrections, deletion and restriction of processing, and the right to appeal. Some of the exceptions are, however, constrained in specific cases. Art. 23 provides further limitations allowed in the GDPR.

The last paragraph of Art. 23 contains a catch-all phrase related to the topic: (3) Where the exclusion or limitation of certain rights or obligations would be likely to result in a high risk to the legitimate interests of the data subject, the controller or processor shall, without undue delay, adopt and document appropriate measures to mitigate such or similar risk. It might be interesting to note that the word *academic* has indeed been used in line with the usual legal text practice, even though outside of the legal domain, it has a meaning very similar to the French expression ‘académique’ (see the discussion below about France), with several tens of “Academic” research institutes formed across the country.

In practice, the formal and often informal guidelines of the Czech Office for personal data protection⁹ are that reporting according to the law should be minimised to clear cases of processing personal data, and only in the case it is not covered by other provisions of the law. For example, if human subjects performing tasks (non-medical) in a research project are being paid by the same institution, their personal data are being collected for processing their salaries and covered by the reporting done once by that institution for the purpose of employment. In such a case, no other reporting is necessary provided the writings, speech recording, survey results or other data collected from the subjects are anonymised (or collected anonymously) before they are stored and processed, which is often the case in linguistic research focused on data collection for machine learning in the area of language technology, where in fact individual differences are to be suppressed anyway to get generalised behaviour of the models and resulting software tools and applications.

The **Estonian** Personal Data Protection Act (Estonian PDPA) has two sections to protect freedom of speech (Section 4 and 5). Section 4 of the Estonian PDPA regulates the processing of personal data for

⁹ Available at <https://www.uoou.cz/en>.

journalistic purposes¹⁰ and is not addressed here. Section 5 concerns the processing of personal data for academic, artistic and literary expression, which is the focus of the article. According to Section 5 of the Estonian PDPA “Personal data may be processed without the consent of the data subject for the purpose of academic, artistic and literary expression, in particular, disclosed if this does not cause excessive damage to the rights of the data subject”. The Explanatory Memorandum to the Estonian PDPA emphasises that the regulation applies *inter alia* to books, motion pictures, visual art, biographies and other content that does not qualify as journalism. According to the memorandum, consent as a legal basis for processing PD for academic, artistic and literary expression is not required. The reason is that consent can be withdrawn, which could have an adverse impact on the freedom of expression. Although the law allows processing PD without consent, it is necessary to strike a fair balance between FoE and privacy (2018: 13-14).

Article 80 of the **French** Data Protection Act attempts to reconcile data protection and freedom of expression in France. It derogates from two general principles: the storage limitation and the prohibition of processing sensitive data (including data about criminal convictions and offences). It also limits information rights, access, rectification and restriction, and derogates from the rules on data transfers. This framework applies only when necessary to safeguard freedom of expression and information, and only when the data are processed: 1) for academic (‘universitaire’), artistic or literary expression, or 2) for journalistic purposes by professional journalists, in a way that respects ethical rules (deontology) of the profession. The Article clearly states that other laws and codes regarding violations of privacy and reputational damage continue to apply.

One can be surprised by the adjective ‘universitaire’ in Article 80 of the French Copyright Act (expression universitaire, artistique ou littéraire) rather than ‘académique’ (as in ‘academic, artistic or literary expression’). However, the same wording is used by the French version of Article 85 of the GDPR. This is because ‘académique’ has a very restricted meaning in French (related to the Académie Française) and should not be interpreted as limiting the derogatory framework to processing made by scholars with a university affiliation.

Article 28 of the **Greek** Personal Data Protection Act, corresponding directly to the GDPR (Art. 85), aims to reconcile the right to personal data protection with the right to freedom of expression and information, “including the processing for journalistic purposes and for purposes of academic, literary or artistic expression”. More specifically, in the framework of these objectives, Paragraph 1 of this Article explicitly enumerates cases where the processing of PD is allowed: “(a) when the subject of the data has given his explicit consent, (b) for PD that have been publicised by the subject, (c) when the right to the freedom of expression and the right to information outweighs the right to PD protection, especially for topics of general interest or when the PD relates to public persons, and (d) when it is restricted to the necessary measure to ensure the right of expression and the right of information, especially with regard to sensitive categories of PD¹¹, and criminal cases, and security-related measures, taking into account the right of the subject to his private and family life.” We can deduce that the Article looks more into the ‘journalistic purposes’ rather than ‘academic purposes’. Paragraph 2 of the same Article provides the exceptions and derogations for processing for such purposes, which are mentioned in Article 85 of the GDPR.

Article 136 of the **Italian** Personal Data Protection Code (PDPC) implements Art. 85 of the GDPR. It regulates journalistic as well as academic works. Article 137 defines the categories of PD that can be processed without the data subject’s consent. Namely, such categories are special categories of PD and

¹⁰ The Estonian Personal Data Protection Act § 4: “Personal data may be processed and disclosed in the media for journalistic purposes without the consent of the data subject, in particular disclosed in the media, if there is public interest therefor and this is in accordance with the principles of journalism ethics. Disclosure of personal data must not cause excessive damage to the rights of any data subjects“.

¹¹ This is really a different approach from the Netherlands. The Dutch variant links Art. 85 GDPR to academic expressions. That leads to legal uncertainty whereas the main principles of the GDPR are oriented towards purposes and legal grounds. The Greek approach seems a clarification of the legal ground of the public interest Art. 6 GDPR. The emphasis on public persons is in line with the case law on freedom of expression. This raises the question to what extent journalistic purposes really are comparable to academic purposes.

PD data related to criminal convictions and offences (GDPR Art. 9, 10). Other sections further restrict these categories to “Safeguards applying to the processing of genetic data, biometric data, and data relating to health” (section 2-f) and “Processing entailing a high risk for the performance of a task carried out in the public interest” (section 2-p). Article 137 (3) provides: “It shall be allowed to process the data concerning circumstances or events that have been made known (communicated/disseminated) either directly by the data subject or on account of the data subject’s public conduct”.

In the **Netherlands** Art. 85 GDPR is implemented in a broad way in article 43 of the *Uitvoeringswet AVG*. The Article speaks of the reconciliation of rights for journalistic and academic expressions. As the Dutch lawmaker explains in the parliamentary discussion (*Memorie van Antwoord UAVG 2017-2018*, 34 851, first chamber) Recital, 153 of the GDPR calls for broad implementation. Yet, still, an assessment of the proportionality has to take place. If needed, transparency requirements and access rights of Chapter 3 GDPR can be disregarded. This raises the question of how researchers should deal with transparency requirements inherent to academic research and verifiability. No exemption is possible from the obligation on data protection by design (Art. 25 GDPR). So, a balancing act and applying data protection principles, like data minimisation, is still mandated. Yet, for instance, when necessary, in urgent cases, based on this derogation, a researcher can refrain from filling in questions about the intended processing, where the institution generally would want this in their shared role of controller, based on the obligation to maintain records of processing activities (article 30 GDPR). This is in line with article 1.6 of the Netherlands Higher education and Research Act, which states that academic freedom is taken into account in universities of the Netherlands. A data protection impact assessment as a method for privacy by design could be used to document the balancing act and design a protocol for verifiability.

4.2.4 Different implementation models and a way forward

Different implementation models raise the question of their potential impact. As a general observation, we would emphasise that Article 85 of the General Data Protection Regulation itself is rather vague. There is a good reason for this. The right to freedom of speech does not have a clear scope. Freedom of speech (also academic freedom of speech) is probably differently defined in different EU countries. This means that what could be protected as academic freedom of speech in Estonian is not necessarily identical to France or Greece. The cultural differences are probably reflected and reinforced in divergent GDPR Art. 85 implementation models that are not limited to laws but also extend to legal practice. Since research is becoming increasingly international, this could be a problem.

Personal data protection and freedom of expression are both human rights. This means that one is not prioritised over another. The critical issue is to strike a fair balance between them, as personal data protection should not affect academic freedom of expression. While openness promotes transparency and accuracy of the research, protective measures promote confidentiality. Both aspects are needed to preserve trust in research among fellow researchers and the general public. The tension might be resolved using the principles on Findability, Accessibility, Interoperability and Reusability of data (FAIR 2016) advocated by the European Commission (COM(2020) 66). In practical terms, it means making the metadata open and providing a clear protocol for accessing the content even if it is not offered openly on the internet.

5 Conclusion

We reached the following preliminary conclusions. Firstly, the data minimisation and the attribution right are not contradictory concepts. The acknowledgement of the author is compatible with the GDPR as the compliance with a legal obligation. The attribution does not concern all personal data but only data that is copyrighted.

Secondly, the access right primarily applies to raw personal data. There is no legal clarity regarding the access to data derived from personal data. The information not containing personal data is not within the scope of the access right (even if the information is derived from personal data). The access right

could be exercised to get access to personal data derived from raw personal data. This is not the case when such access could conflict with intellectual property rights and trade secret protection. Trade secret protection considerations could be more relevant here.

Thirdly, personal data protection usually does not take precedence over the freedom of expression and cannot hinder the academic freedom of speech and the author's right to disseminate his work. However, there could be restrictions on how it can be disseminated. Conducting research may also be covered by academic freedom of speech. Although the General Data Protection Regulation provides a framework to enhance freedom of speech in the field of academic research, its implementation by different EU countries diverges. It is not necessarily compatible with the GDPR's aim to establish a uniform framework for processing personal data for research and academic expression and could have a negative impact on the dissemination of research results. Therefore it is advisable to rely on the principles of Findability, Accessibility, Interoperability and Reusability of data.

References

- Austrian Data Protection Amendment Act. Entry into force 2018. Available at <https://www.ris.bka.gv.at/eli/bgbl/I/2018/31> (3.4.2021).
- Berne Convention. *Berne Convention for the Protection of Literary and Artistic Works of September 9, 1886*. Available at <https://wipolex.wipo.int/en/text/283698> (9.2.2022).
- Charter of Fundamental Rights of the European Union. - OJ C 326, 26.10.2012, p. 391-407. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012P%2FTXT> (4.2.2022).
- Convention Establishing WIPO. Convention Establishing the World Intellectual Property Organization (as amended on September 28, 1979). Available at <https://wipolex.wipo.int/en/text/283854> (8.2.2022).
- Convention for the Protection of Human Rights and Fundamental Freedoms. Rome, 4.XI.1950. Available at https://www.echr.coe.int/documents/convention_eng.pdf (4.2.2022).
- Czech Republic Law No. 110/2019 Coll. Available at <https://www.zakonyprolidi.cz/cs/2019-110> (4.2.2022).
- C-434/16. *Case C-434/16*. Peter Nowak v Data Protection Commissioner (20 December 2017). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62016CJ0434&qid=1617649690306> (5.4.2021).
- C-141/12. *Joined Cases C-141/12 and C-372/12*. YS (C-141/12) vs Minister voor Immigratie, Integratie en Asiel, and Minister voor Immigratie, Integratie en Asiel (C-372/12) v M, S (17 July 2014). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0141&qid=1617633666683> (5.4.2021).
- C-145/10. *Case C-145/10*. Eva-Maria Painer vs Standard VerlagsGmbH and Others (1 December 2011). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62010CJ0145&qid=1618044850444> (10.4.2021).
- C-5/08. *Case C-5/08*. Infopaq International A/S vs Danske Dagblades Forening (16 July 2009). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005> (10.4.2021).
- COM(2020) 66. *A European strategy for data*. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels, 19.2.2020. COM(2020) 66 final. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN> (15.1.2022).
- Deutscher Bundestag, 2018. *Deutscher Bundestag*, Ausarbeitung: Die Öffnungsklausel des Art. 85 der Datenschutz-Grundverordnung, WD 3 - 3000 - 123/18. Available at <https://www.bundestag.de/resource/blob/560944/956f5930221c807984d40c1df2af5abf/WD-3-123-18-pdf-data.pdf> (26.04.2021).
- DSM Directive. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *OJ L 130*, 17.5.2019, pp. 92-125. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790> (10.4.2021).
- Dutch Implementation Act General Data Processing Regulation, In force from: 16th of May 2018, available at: Uitvoeringswet Algemene verordening gegevensbescherming, Permanent link to Artikel 43 Uitvoeringswet

- Algemene verordening gegevensbescherming. Available at <https://wetten.overheid.nl/BWBR0040940/2021-07-01> (20-1-2022)
- EDPB 2022. *European Data Protection Board*. Guidelines 01/2022 on data subject rights - Right of access. Version 1.0 (for public consultation). Adopted on 18 January 2022. Available at https://edpb.europa.eu/system/files/2022-01/edpb_guidelines_012022_right-of-access_0.pdf (1.2.2022)
- EDPB 2019. *European Data Protection Board*. Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. Version 2.0. Adopted on 20 October 2020. Available at https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf (8.2.2022).
- Estonian Copyright Act. *Copyright Act*. Entry into force 12.12.1992. Available at <https://www.riigiteataja.ee/en/eli/504032021006/consolide> (9.2.2022).
- Estonian PDPA. *Personal Data Protection Act*. In force from: 15.01.2019. Available at <https://www.riigiteataja.ee/en/eli/523012019001/consolide> (6.2.2022).
- 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> (1.4.2021). FAIR 2016. *Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0*. Available at <https://www.force11.org/fairprinciples> (15.1.2022).
- Explanatory Memorandum to the Estonian PDPA (2018). Available at <https://www.riigikogu.ee/tegevus/eelnoud/eelnou/5c9f8086-b465-4067-841e-41e7df3b95af> (7.2.2022).
- Finnish PDPA. *Data Protection Act (1050/2018)*. Available at <https://www.finlex.fi/en/laki/kaanokset/2018/en20181050.pdf> (8.4.2021).
- French Data Protection Act. *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. Available at <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460/> (27.4.2021).
- French Intellectual Property Code. *Code de la propriété intellectuelle*. Available at https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006069414/ (27.4.2021).
- German Federal Data Protection Act. *Bundesdatenschutzgesetz vom 30. Juni 2017 (BGBl. I S. 2097)*, das durch Artikel 12 des Gesetzes vom 20. November 2019 (BGBl. I S. 1626) geändert worden ist. Available at: https://www.gesetze-im-internet.de/bdsg_2018/BJNR209710017.html (27.4.2021).
- 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> (1.4.2021).
- Greek Data Protection Act. *Personal Data Protection Law (4624/2019)*. Available at: <https://www.e-nomothesia.gr/kat-dedomena-prosopikou-kharaktera/nomos-4624-2019-phek-137a-29-8-2019.html> (26.4.2021).
- InfoSoc Directive. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal L 167*, 22/06/2001 P. 0010 – 0019. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029> (4.4.2021).
- Italian PDPA. *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.garanteprivacy.it/documents/10160/0/Data+Protezione+Code.pdf/7f4dc718-98e4-1af5-fb44-16a313f4e70f?version=1.3> (8.4.2021).
- Kamocki, Pawel. 2020. When Size Matters. Legal Perspective(s) on N-grams. *Proceedings of CLARIN Annual Conference 2020. 05 – 07 October 2020*. Virtual Edition. Ed. Costanza Navarretta, Maria Eskevich. CLARIN, 166-169. Available at https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf (10.4.2021).
- Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramunas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värvi, Pavel Stranák, Jan Hajic. 2020. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. In: Kiril Simov, Maria Eskevich (Ed.). *Selected*

- Papers from the CLARIN Annual Conference 2019 (53–65)*. Linköping University Electronic Press. Available at <https://ep.liu.se/ecp/172/008/ecp20172008.pdf> (10.4.2021).
- Kelli, Aleksei, Tõnis Mets, Lars Jonsson, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Age Värv. 2018. Challenges of Transformation of Research Data into Open Data: the Perspective of Social Sciences and Humanities. *International Journal of Technology Management and Sustainable Development*, 17 (3), 227–251.
- Latvian PDPA. *Latvian Personal Data Processing Law*. Available at <https://vvc.gov.lv/image/catalog/dokumenti/Personal%20Data%20Processing%20Law.doc> (26.4.2021).
- Lithuania PDPA. *The Republic of Lithuania Law on Legal Protection of Personal Data*. Available at <https://www.e-tar.lt/portal/legalAct.html?documentId=43cddd8084cc11e8ae2bfd1913d66d57> (26.4.2021).
- Netherlands Higher education and Research Act, In force from: 8th of October 1992, Available at: <https://wetten.overheid.nl/BWBR0005682> (7.2.2022).
- Trade secrets directive = Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure. – OJ L 157, 15.6.2016, p. 1-18. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016L0943> (4.2.2022).
- Universal Declaration of Human Rights. Available at <https://www.un.org/en/about-us/universal-declaration-of-human-rights> (4.2.2022).
- WP29 2016. *Article 29 Working Party (WP29)*. Guidelines on the right to data portability. Adopted on 13 December 2016. Available at https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf (5.4.2021).
- WP29 2007. *Article 29 Working Party (WP29)*. Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (5.4.2021).

Collaborating on Language Resource Infrastructures with Non-Research Partners: Practicalities and Challenges

Verena Lyding

Eurac Research, Italy
{verena.lyding, egon.stemle}@eurac.edu

Egon Stemle

Eurac Research, Italy
{verena.lyding, egon.stemle}@eurac.edu

Alexander König

CLARIN ERIC, The Netherlands
alex@clarin.eu

Abstract

By now, digital infrastructures for language data and tools have become commonplace in the research domain, but their possible benefits are still almost unknown outside of these circles. However, it stands to reason that the data and methods developed there could also be used by non-research language actors like publishing houses or libraries. This article presents a use case within a local language infrastructure project describing our interactions with a newspaper portal that resulted in modern NLP tools being made available via an API to help improve their online search. We describe how this use case was implemented, focusing on the problems that came up, specifically those from the interaction between a research and a non-research institution.

1 Introduction

Large scale research infrastructure projects like CLARIN (De Jong et al., 2018), DARIAH (Edmond et al., 2017) or ELG (Rehm et al., 2021) aim at making language resources and tools available to, sustainable for and easily reusable by their stakeholders. These efforts have proven to create standards and frameworks and have become a reference point for visibility. Yet, up to today, the active involvement of stakeholders and the ambition to attract users to the provided services and tools is challenging. For example, different User Involvement (UI) events of CLARIN helped to provide specific training to a number of research stakeholders¹ and the CLARIN Resource Families initiative (Fišer et al., 2018) links resources of several research stakeholders. Still, stakeholders from industry are hardly found among the users, and experiences from projects that actively involve commercial partners show that the industrial use of the offered services is indeed difficult.²

This is related to the naturally slow advancement of large scale, complex and usually abstract projects. In particular, solutions that aim at encompassing various use cases and demands tend to result in powerful yet generic frameworks, as for example, the Component Metadata Infrastructure³ (Goosen et al., 2014). Those solutions are not always easy to adopt because they require knowledge and technical skills, and it is not always clear from the onset whether the actual use case can be implemented. For this reason, and to bridge the gap between research and application, projects are created that cover domain-specific use cases with the help of large infrastructures (for example, ELEXIS (Woldrich et al., 2021)).

This paper presents a use case from the local language infrastructure project DI-ÖSS⁴(Lyding et al., 2019). This project bridges the gap between an existing infrastructure project (CLARIN) and a local community. That is, instead of targeting a specific application *domain* (like e.g. lexicography) DI-ÖSS targets a wider set of *local stakeholders* that are working with language in different ways. By doing so DI-ÖSS aims to connect those local actors to ideas, procedures and solutions from the large infrastructure.

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

¹See, for example, <https://cmc-corpora2017.eurac.edu/uievent/> and an overview here: <https://www.clarin.eu/content/user-involvement-funding#guest-blog-posts>

²Bleichner et al. (2005) report on a cooperation between two German universities and a part of the archiving division of AIRBUS; Poesio and Magnini (2009) report on a project with data providers of audio, video and text news from Trentino, Italy.

³<https://www.clarin.eu/cmdl>

⁴*Digitale Infrastruktur für das Ökosystem Südtiroler Sprachdaten und -dienste* - Digital infrastructure for the ecosystem of South Tyrolean language data and services

2 Background

2.1 Local Infrastructure Project DI-ÖSS

The presented use case on an interaction between an online newspaper portal and an NLP service hosted at a research institution was carried out as part of the small local infrastructure project named DI-ÖSS (Lyding et al., 2019) which ran from 2017 to 2021. The aim of the DI-ÖSS project was to connect various types of language actors on the local level to exploit synergies between their activities and goals and the objectives of Eurac Research's Institute for Applied Linguistics (IAL). The IAL is a member of CLARIN-IT and is the initiator and leader of the DI-ÖSS project. The project explicitly aimed at the involvement of non-research partners, which are typically not familiar with infrastructure efforts on the European level. A consortium with four local language actors was established to explore different use cases. Next to the newspaper portal two public cultural institutions, a local library partner and a public culture and language institution, as well as a non-computational research partner working with historical letters were part of the consortium. The model of cooperation between each of the four local language actors and the IAL was an asymmetric project cooperation, with the major workload on the side of the IAL, and a smaller workload on the side of each of the partners, limited to accompanying the use case development with their relevant institutional knowledge. Any active data curation and development work was delegated to subcontractors, which were coordinated by the IAL and paid by the project budget.

2.2 Finding Partners

The first phase of the project consisted of the lead partner at IAL searching for cooperation partners. These partners should be outside of the area of research as the aim of the project was to widen the idea of *infrastructure for language data* beyond the scope of research where it is well established by now. As a first step an extended list was created of institutions that primarily work with language data within the project's dedicated geographical region, the Autonomous Province of Bolzano/Bozen in northern Italy. There can be a case made that any institution is dealing with language data to some degree, but for this list the focus was on institutions where the language data is their main focus. In the end, this list contained about 200 entries that could be grouped roughly into seven categories: archives, libraries, online media, catalogs, language units, publishing houses and journals. See Figure 1 for their distribution.

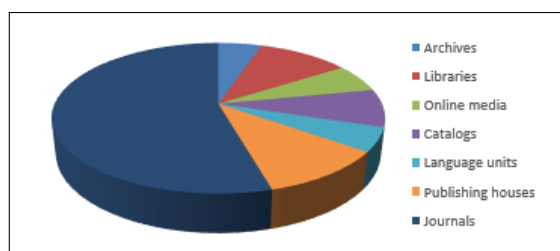


Figure 1: Distribution of institution types in preliminary classification

The overabundance of journals in our long list is due to the fact that they are very visible carriers of local language data, even though a lot of them are not produced by what can be called a "language institution". We kept them in the list, because they can be a rich source of local language data, but drastically reduced their weight in the following steps, namely the interview phase described below.

The IAL did in-depth interviews with eleven individual institutions, trying to cover all of the various categories and also looking at institutions of different sizes. These interviews were used to get some idea of typical processes within these institutions (Lyding et al., 2020) and starting from that develop some possible use cases that could be worked on with such an institution within the scope of the project.

The IAL then invited some institutions as cooperation partners into the DI-ÖSS project. As during the interviews, it was tried to cover a wide range of different institutions. But it turned out to be surprisingly difficult to convince non-research institutions to join this kind of project. We encountered quite strong reservations regarding whether participating in such a project would be worth the institution's time. We assume that this can be explained by the fact that especially institutions organised as a business have to

always calculate whether the time (and therefore money) they invest into such a new project will be met by enough revenue, that is, will they get enough out of it or even more general *what* can they get out of it. For partners from the newspaper publishing world, three out of the four institutions we contacted were not interested in any cooperation even though, apart from the possible technological benefits, a small monetary reward was offered.

2.3 Defining Use Cases

Another problem that paired with this general reservation was the challenging task of envisioning possible use cases that could be explored in such a cooperation. When setting up the cooperation with the newspaper portal, we tried to develop a use case together with the people working there. But we realised that it was difficult for them to see beyond their day-to-day work within an established environment and come up with ideas that could utilise the possibilities lying in such a cooperation with a research institution. This showed that it is difficult for a potential industrial partner to envision a possible use case using NLP tools and other language technology methods because the extent of these methods is not widely known. Therefore, users either have no idea at all what is possible or on the other hand greatly overestimate the power of these tools and come up with ideas that are virtually impossible with today's capabilities. Also, the factor of having to adapt established workflows can pose obstacles for businesses offering professional services. Any adaptation can lead to a possible disruption of a workflow, and it is therefore understandable that non-research partners are particularly wary of committing to 'unnecessary' changes to a running system, even more so if the added value is something abstract like an evaluation metric, a promise of an improved experience, or a functioning prototype with different data or not fully integrated into their usual workflow. While we can envision potential use cases and the expected added value of a project, the cooperation with a research partner and research tools cannot be guaranteed to be as stable and predictable as commercial services. Insofar, it was a fine line we had to walk in order to entice partners with possible opportunities on the one hand, but also not to promise too much.

2.4 Related Work

The research community is in active exchange on language research infrastructure initiatives as conferences by the main players CLARIN, META-SHARE and ELG, and dedicated conference tracks at NLP conferences⁵ show. Also, calls for application showcases are widely promoted and fostered with financial incentives^{6,7}. Despite this active promotion of the adoption of language research infrastructures by a wider audience, it is extremely difficult to find scientific relations and reports on research - industry cooperations. This even holds for the inter-institutional projects mentioned above.

While it would be very valuable to gain insights on prior experiences with adopting research infrastructure components for use cases from industry, the lack of these types of publications is also comprehensible. Research - industry cooperations are challenging by nature and in the context of project-based initiatives often experimental and small in scale. If achieved, results may remain preliminary and use cases might not always turn out as success stories, thus the motivation to publish about it can be naturally diminished. In addition, scientific conferences generally target substantial scientific contributions and concluded works rather than work-in-progress reports. In conclusion, our search for related works has been without noteworthy results and it remains to the scientific community to encourage more project reporting on less shiny but insightful use cases and cooperations over time.

3 Cooperation with a Newspaper Portal

The use case explored further in this paper is built on the inter-institutional cooperation among the Institute for Applied Linguistics (IAL) at Eurac Research⁸ and the local newspaper portal [salto.bz](https://www.salto.bz)⁹. Among the cooperations with the four project partners (see above), we decided to focus on this single cooperation

⁵<https://lrec2022.lrec-conf.org/en/calls-papers/2nd-call-papers/>

⁶<https://www.clarin.eu/content/user-involvement-funding>

⁷<https://www.european-language-grid.eu/open-calls/>

⁸<https://www.eurac.edu/linguistics>

⁹<https://www.salto.bz>

as the local newspaper portal was the most commercial/industry partner within the consortium. Moreover, the local newspaper was the only partner we had not worked with before. Accordingly, this collaboration was the most challenging and insightful in terms of identifying a use case and implementing it.

3.1 Cooperation Partner: Salto.bz Newspaper Portal

Salto.bz is a 'news and community portal for South Tyrol'. It has been founded in 2012 as a cooperative society and its news portal is online since 2013.¹⁰ Salto.bz is the first German and Italian bilingual online news portal in the multilingual province of South Tyrol. It was created with the aim to combine journalism and social media communication and offers editorial content of professional salto.bz authors as well as texts, comments and multimedia content provided by its community. It focuses on journalism and information exchange on daily news and analyses on politics, economy, environment and society.

Among the different news publishers in South Tyrol salto.bz stood out by its openness, interest and availability for a cooperation with us as research institution. In contrast to experiences from multiple other attempts to find collaboration partners among local publishers the head office of salto.bz immediately signalled willingness to learn more about the project idea, to discuss specific cooperation possibilities, and to promote the cooperation to its internal workers and to involve them where relevant.

3.2 Use Case: Advanced Search

The initial discussions between the IAL and salto.bz focussed on understanding the structure of the newspaper portal, the uses-cases of their internet users, the types of data which are created and administered by salto.bz and the related workflows from article writing to publishing. The aim was to identify a use case that could benefit from NLP treatment. This way, the IAL could offer an NLP service to salto.bz while, in return, getting access to authentic language data produced in the bilingual context of South Tyrol to carry out linguistic studies.

Analyzing internal workflows at salto.bz, the following things could be observed. Concerning the back-end interaction, it became clear that authors do not perform any language processing activities within the portal. They mainly interact with it to upload and publish new articles. In addition, news writing and publishing is often carried out under time pressure, and any additional activity required before completion is not appreciated by authors, unless the added value is very clear. Concerning the front-end interaction, that is the interaction of readers with the portal, the search functionality within current news and news in the archive showed some shortcomings in terms of search speed, support for multiword searches and search by criteria like publication date, author or ressort.

The use case we jointly identified targets the creation of an improved search service for the articles of the news portal, including recent articles and the entire news archive. To be most useful to salto.bz users, the outline of the improved search service foresaw the following functionalities:

1. Full-text search with support for multiword searches,
2. Facetting by manually created metadata for each article (i.e., author name, publication date and section of the newspaper), and
3. Automatically generated keywords as additional facets to refine the search results.

Thus, the use case implements a service of general interest for the news portal and includes an NLP component prototype which is delivered by the IAL as research partner. The distribution of the work between salto.bz and the IAL of Eurac Research was organised as follows: salto.bz took care of the general technical aspects of the portal programming, while the IAL provided a computational linguistic component for multilingual keyword extraction from news articles. Detailed information on the distribution of work and the interaction between project members across institutions are given in Section 4.3 below.

¹⁰<https://www.salto.bz/de/faq>

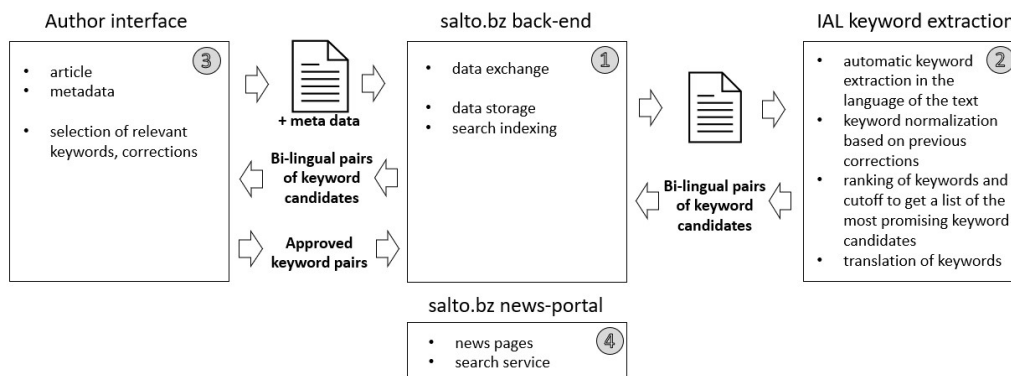


Figure 2: System architecture

3.3 Coordination of the Inter-Institutional Interaction

The inter-institutional cooperation was initiated by the project coordinator of the IAL who contacted the salto.bz head office and proposed a project cooperation. The salto.bz head office agreed to be part of the project and participated in several meetings with the IAL to discuss the working reality and content on the side of the newspaper portal and to identify the specific use case for the cooperation. Once the use case was established, on salto.bz side the overall interaction continued to be handled by the head of the back office, who administered contractual aspects and coordinated the participation of salto.bz employees in the use case, mainly the interaction with the programmer of the portal. On IAL side the cooperation was coordinated by the project lead in consultation with its other researchers. The technical implementation was carried out between the programmer at salto.bz and the project researchers at the IAL. As soon as the system components were in place, also the editor-in-chief of salto.bz got involved on the client side to give feedback on the user interaction within the author's interface. The editor-in-chief also took care of the communication with the authors of salto.bz. Finally, the IAL involved a translation expert to evaluate and curate the automatically extracted keyword pairs.

4 Implementation of the Improved Search Service

4.1 System Design

The extended search service is designed as a distributed architecture with the search interface running on the salto.bz news portal and the computational linguistics text processing being performed at the IAL.

Overall, we can distinguish four components of the system architecture (see also Figure 2):

1. The portal back-end, data storage and search engine (salto.bz¹¹)
2. The keyword extraction service (IAL)
3. The author web interface (salto.bz)
4. The news search web interface (salto.bz).

The portal is centrally based on Drupal¹², an open-source content management system (CMS) that can be extended with modules to expand its functionality. Thus, the functionality on the salto.bz-portal side was fully integrated into the CMS and its regular workflow. This means the authors enter their articles via a web interface and optionally activate a *Get Tags*-function, which triggers a process on the back-end side

¹¹Strictly speaking, the division here is subdivided again: The CPU time, the data storage, the general Drupal and Apache Solr/Lucene installations are provided by an external Internet service provider.

¹²<https://drupal.org>

of the CMS to send text with metadata to the keyword extraction service to retrieve candidate keyword pairs. The keyword extraction service at the IAL receives data from the portal back-end, processes the news article texts and metadata, identifies keywords, sends new keywords for translation to an external service, and returns candidate keyword pairs back to the portal back-end. These candidates are then forwarded to the author’s web interface for validation. The author’s interface allows to validate, delete or modify candidate keyword pairs before the article is queued for review or publication.

4.1.1 The News Search Web Interface

The news search is based on the integration of Drupal with Apache Solr¹³, a popular open source enterprise search platform built on Apache Lucene. The web interface allows for full text search and faceting of results both by metadata information and keywords. Figure 3 shows the results page for the search *Fahrrad* (*Bicycle*). The articles that match the query are shown on the right, with their section (e.g. *UMWELT* (*Environment*)), title, author, publication date, and highlighted text for the query match(es). The left side (or a popover on smaller screens) shows some of the possible facets to restrict the search

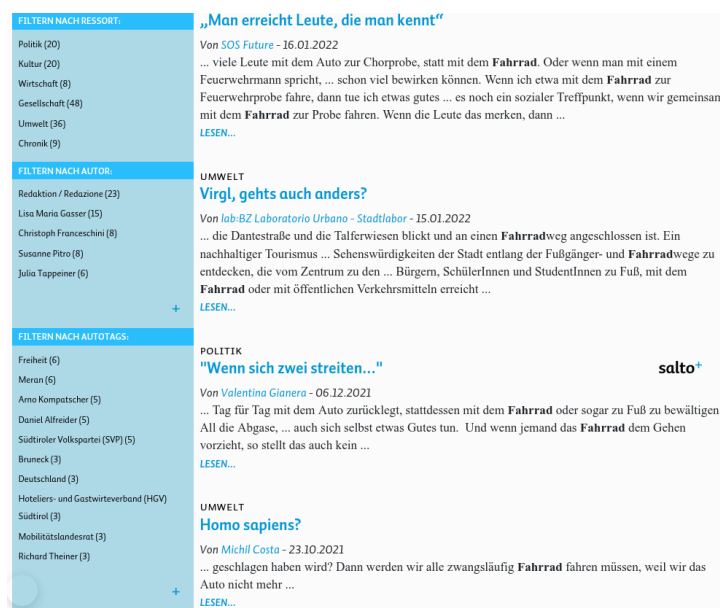


Figure 3: Search results with (some) facets

results: *FILTERN NACH* (*Filter by*) *RESSORT* (*Section*), *AUTOR* (*Author*), and the generated *AUTO-TAGS* (*keywords*). The first facet, which is at the beginning of the list and not visible any more on this scrolled down view, is *year of publication*. Further down the facets list is also a short info box containing information about search in general and the Autotags feature (our translation):

A search for several words is possible. For an exact search, the terms can be placed in inverted commas.

The number of hits per category is displayed in the search options. Clicking on a value filters by category.

The category 'Autotags' shows keywords that have been automatically generated using computer linguistic methods. This functionality is the result of a research cooperation with Eurac Research and is currently in beta phase.

4.1.2 The Author Web Interface

The authors of salto.bz enter news content through the author web interface which provides a form with distinct fields for the news body text, title, section, etc. In terms of content creation, news articles and metadata such as section or publishing date are manually created by the authors of salto.bz, while keyword pairs for each text are generated on demand by the keyword extraction service of the IAL. After entering the news text, authors have to actively retrieve and validate keyword pairs through the system. Figure 4 shows the author web interface with the part for the keyword extraction highlighted. Keywords

¹³<https://solr.apache.org/>

are generated by clicking on the 'Get Tags' button. The automatically extracted keywords will appear below the text field; light grey indicates new keywords and turquoise keywords that already exists in the system. Suggestions can be approved (\oplus), which adds them to the text field, or can be cleared from the text field (\ominus). New keywords can also be freely added to the text field or any approved one can be changed.

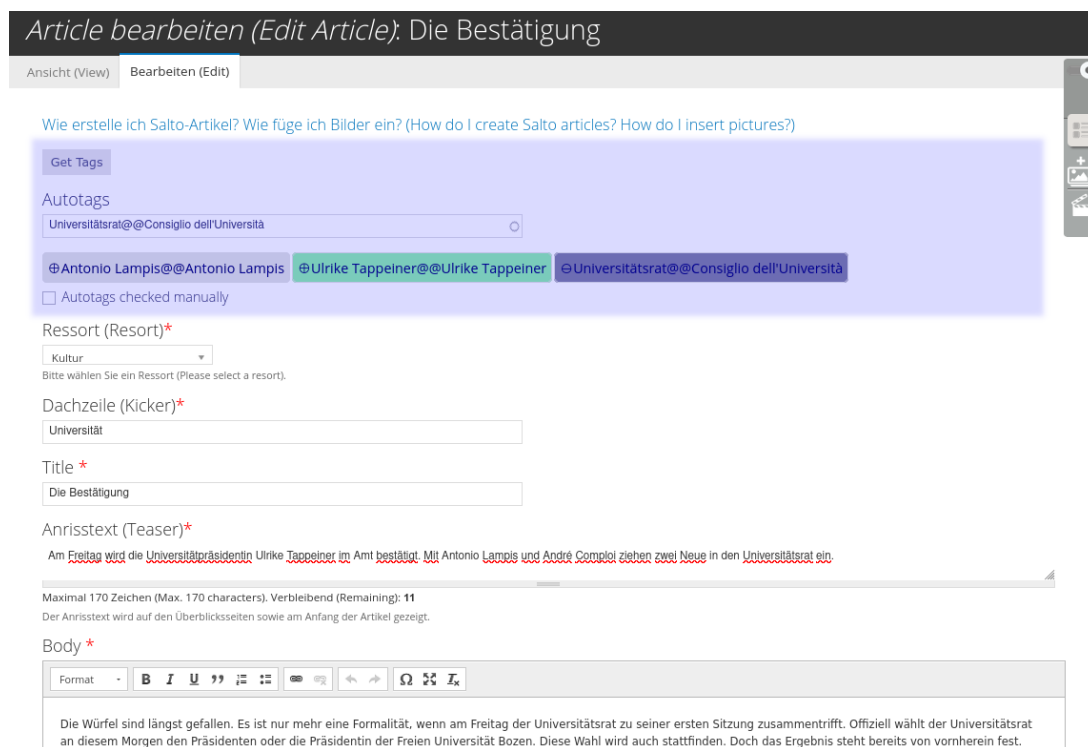


Figure 4: Author's web interface for editing an article and the keyword extraction 'Autotags' highlighted

4.2 Technical Implementation of the Keyword Extraction Service

The keyword extraction tool is a prototype implementation which extracts keyword from an article and uses the Microsoft Bing translation API¹⁴ to translate each extracted keyword into the respective other language. That is, it receives a text in either Italian or German and returns a list of bilingual keyword pairs in the form *German@@Italian*. Together with the developer of the newspaper portal, we defined three requirements for user interaction. The interface should allow (1) to retrieve a list of candidate keyword pairs on demand, (2) to select or deselect candidates according to their relevance and (3) to change or correct the selected candidates if necessary.

Keywords for us are single or multi-word expressions that do not necessarily have to occur in sequence in the text and are extracted by a handful of manually devised rules¹⁵. Through the rules, we were able to ensure that some peculiarities of typical local naming, for example, names of public administration entities, can be recognised in both German and Italian. The translation was done for each keyword independently of its context - this is a borderline use of the translation service, but for a working prototype we accepted this shortcoming. This implicates that for ambiguous keywords the correct translations cannot be guaranteed by the system, but are enforced in the manual validation step instead.

On the technical side, the exchange between the parties was standardised. All involved parties could develop their system independently, and still, the systems could communicate with each other. To this end, a RESTful application programming interface (REST API) (Fielding, R. T., 2000) was designed and implemented that allows documents to be sent to the IAL for computer-aided linguistic processing, which

¹⁴<https://docs.microsoft.com/en-us/azure/cognitive-services/translator/>

¹⁵See `keyword_extractor_salto.py` in <https://gitlab.inf.unibz.it/commul/di-oss/api-service-salto>

can be retrieved after successful processing. The API implements an authentication and authorisation layer that allows us to track which document came from whom so that different processing or licensing agreements can be considered.

The processing of the documents is (optionally) asynchronous in order to be able to take into account long-lasting processing. For this purpose, the API assigns an identification token after successfully transmitting a job, which the remote party must use when returning to query whether processing has already been completed. Once processing is complete, the bilingual keyword pairs suggested by the system can be retrieved and are displayed to authors for further processing as described above.

The keyword pair candidates are compared with those already in the system, and the already known pairs are colour-coded. As explained above, the suggestions must be actively accepted, i.e. there is no mechanism that automatically assigns the suggestions to an article. In addition, the provision of keyword candidates is beneficial but not critical. In the absence of such suggestions, authors can also complete editing an article without automatic suggestions by manually entering their suggestions, which are automatically completed with the suggestions known to the system.

Keyword pairs can also be curated in a separate interface. Changes in this interface are recorded so that traceable changes can be systematically automated. For example, a singular-plural association can be made, which is then recorded as an entry in a file and considered for future proposals. In this way, an author could benefit from both the automated system (the proposal) and the regular maintenance of the taxonomy in a future article. In order to ensure an exchange of information about the acceptance or the content of the accepted keywords, a data reconciliation is carried out at regular intervals. For this purpose, the log file is provided by the developer of the news portal and transferred to our system.

It is important to underline that the keyword extraction service is a prototype implementation not a final product. It has been implemented for the purpose of getting a viable use case up and running within a project cooperation between a research and a business partner. Given this restricted scope and related time constraints, no formal user evaluation (neither with news authors nor with the portal users) has been carried out, but informal feedback on the service has been collected from the news authors throughout the project and indicated that the overall keyword quality was considered acceptable.

4.3 Interaction between Project Members across Institutions

In addition to the overall coordination of the project work across the two institutions (cf. Section 3.3) in particular the implementation work had to be orchestrated between the IAL and salto.bz and the different roles of the project participants. The technical implementation of the different system components were divided as follows: salto.bz was in charge of implementing the back-end of the full text search with faceting by metadata, as well as the user search interface and the integration of the keyword extraction tool within the author's interface. The IAL was in charge of implementing the keyword extraction tool and making it available as an independent service. While the implementation of the components were handled independently by the technical profiles (developer/researcher) at both institutions, several interactions were needed to define technical and user-related requirements. The project core team, composed of researchers at the IAL and the salto.bz developer, regularly met to define which functionalities the search interface and the author interface should include and how they should be presented to users of the portal and newspaper authors. The resulting design specification were then first presented to the salto.bz head office and after to the editor-in-chief to collect feedback for the search and author interfaces. After the details of the system design were set, the core team worked on defining the data exchange formats and protocols, which served as basis for the widely independent implementation of the different components by salto.bz and the IAL. Therefore, once the first version of the entire system architecture was put together again the salto.bz head office and the editor-in-chief were involved for system testing from the users' perspective.

Apart from overall design decisions, the cooperative work on this use case mainly concerned the interaction between the author interface of the news portal and the keyword extraction service offered by the IAL. Both with regards to the technical data flow and the user interaction, decisions had to be taken about the number and order of keyword pair candidates, and about how to select and correct them.

4.4 Required Manual Input of News Authors

The system design, as described above, included the authors' keyword validation and curation activity as a fundamental part of the entire search service. Given that the automatic keyword extraction tool can only work so well in extracting relevant keywords and proposing valid translations for the given context, the project team decided that the manual validation of each keyword pair would be the base condition for including keywords as facets within the public search interface.

Once the new search service with keyword extraction was running in beta-version, all salto.bz authors had to be brought on board. The authors had to be made aware of their task to generate (launch the automated process) and approve (manually select, discard and/or correct) relevant keyword pairs. This required teaching them how to use the tool and motivating them to use it for every article they write. To motivate the authors, the added value of the keyword tool was communicated (the keywords serve as additional search facets), and the editor-in-chief encouraged participation.

The project core team provided guidelines for selecting or correcting keyword candidates. Finally, authors were also instructed to document troubles and errors they encountered, as a functioning feedback loop is essential to maintain and improve the service. In fact, we encountered several situations where authors had noticed an error and stopped using the tool without informing us.

4.5 Updating Existing Data

Given that the assignment of keyword pairs to articles requires human approval, by the time the search interface should go live, we had to find a solution for articles written and published before the keyword extraction tool had been introduced. Asking authors to manually validate keyword pairs for all articles in the archive was not feasible, therefore we had to find a way to update the articles from the archive in an automatic way. We proceeded by applying the automatic keyword tool to all articles of the archive, but kept only those keywords that occurred more than ten times, while discarding all the others. This resulted in a list of more than 2000 keyword pairs, which were manually corrected (e.g. for translation errors) and merged by a translation expert, who was subcontracted for this specific task.

The merging task that is the matching of keyword pairs which are not identical but refer to the same entity or concept was relevant for the keyword list of the articles in the news archive, but is expected to become recurrently relevant also in the future. Given that every day new keyword pairs are generated for new articles, it is likely that variants or semantically similar keyword pairs (almost synonyms) will accumulate and will need to be merged manually to keep the overall amount of keywords under control.

5 Evaluation of the Inter-Institutional Cooperation on Language Infrastructures

The primary goal of the DI-ÖSS project, and more specifically on the cooperation between the IAL and the salto.bz news portal, was to explore as many aspects of an inter-institutional cooperation on language infrastructures as possible. In this respect, the presented use case is to be understood as an all-encompassing feasibility study and not primarily as a technical one. This also means that creating a functioning prototype was one partial and practical aspect to exercise the inter-institutional collaboration on a practical use case. The use case of creating an extended search service, in the first place focused on creating and trial running a workflow that allows for the inter-institutional exchange and processing of texts, their integration into a running newsportal and the interaction of automatically applied procedures (the keyword extraction) with manual processing and validation tasks by the news authors.

In the following subsections, we will present a short evaluation of four relevant aspects of the inter-institutional collaboration: (1) technical interaction and performance, (2) quality of automatically and manually processed data, (3) added value for both institutions, and (4) sustainability of the cooperation.

5.1 Technical Interaction and Performance

On the technical level, the interaction was implemented as a RESTful API without major obstacles, however, a number of factors had to be considered and taken care of: (1) security protocols, (2) monitoring and error messages/reporting, (3) time delay, and (4) updates of services, program versions, etc.

example	error type	correction strategy
<i>Ich bin mit Bozen@@Io sto con Bolzano</i>	entire phrase	limit # of (function) words
<i>Alexander Huber@@Alessandro Huber</i>	translation of person names	list of personal names
<i>Flüchtling@@profugo</i>	differing singular plural conventions	manual correction
<i>Kandidatenliste@@Candidati</i>	translation compounds to MWEs	dictionary look-up

Table 1: Error types of keyword pair candidates

Several interaction steps had to be established for the exchange between the authoring interface of the news portal and the language service provided by our keyword extraction and translation tool.

5.2 Quality of Automatically and Manually Processed Data

The quality of automatically generated candidates of bilingual keyword pairs had to be assessed on two levels: On the one hand the keyword pair has to be formally correct and meaningful in itself, and on the other hand the keyword pair has to make up for a meaningful label of the given text. Throughout development we manually evaluated the generated keyword pairs and identified a number of recurrent error types, such as longer phrases or translations of person names. Table 1 lists common error types and strategies applied for their correction.

On the NLP side, we are aware that keyword extraction and translation are in themselves extensive topics within computational linguistics, but ultimately we opted for pragmatic solutions to get the use case up and running within a restricted time frame. This was also made all the easier by the fact that it was desired from the side of the content creators (the news authors) to have full control over all automatically generated output by being able to check and change the automatically generated keywords as 'suggestions' individually. The quality of the suggestions was one important aspect, but the integration of automatic methods and manual quality control within one workflow was the primary one.

During beta testing the adequacy of keyword pairs for news articles was assessed both by the authors of salto.bz as well as by a translation expert, specifically appointed for the quality evaluation of the bilingual keyword data. While a larger number of problems was identified with the formal correctness of keyword pairs as explained above, fewer issues were encountered with their adequacy to describe the article's context. Indeed, most cases of inadequate keyword pairs related to keywords that fit the text, but are little meaningful as keywords, such as *Datum@@data* (engl. date), *klein@@piccolo* (engl. small), *Michl@@Michl* (a first name) or *Nein@@No* (engl. no).

In addition, we encountered a number of keywords with similar semantics that should eventually be merged into one keyword, (i.e., keywords referring to the same concept such as *offener Brief@@lettera aperta* (engl. open letter) and *Brief@@lettera* (engl. letter) should be merged). Having several keywords for one concept is particularly unfavourable in the context of the portal archive search, for which the generated keywords are used as a search facet and apparent 'duplicates' should be avoided.

Since new keyword pairs get added to new articles over time, the coherence of the overall set of keywords needs checking and merging at regular intervals. As an activity that is asynchronous to keyword assignments to single articles it needs particular attention and detailed knowledge of the database and existing keyword pairs. Therefore, this task is best carried out by a professional with dedicated time for it. Also, the correction of keyword pairs turned out to be difficult for some authors, who might not be fully bilingual. To solve the translation and harmonisation issues, we hired a translation professional that carried out the merging and correction task on a weekly basis during the trial phase.

Since the portal search is publicly accessible by all salto.bz costumers, the keyword results need to live up to a minimum quality standard, which can only be guaranteed with regular curation, and indeed the expert role performing the curation would be needed throughout time, which poses a considerable demand for the sustainability of the use case (see Section 5.4 below).

5.3 Added Value for Both Institutions

As the idea of the DI-ÖSS project was to develop use cases to show the potential synergies of establishing a local small-scale infrastructure with non-research institutions one important measure of success was in how far the work that has been done provided added value for the stakeholders involved. For the IAL the added value is very real. Every time an article is sent to the keyword generator, we save a copy of it to our salto.bz corpus that can be used for linguistic analyses, both of Italian and South Tyrolean German.

For salto.bz there are two groups that have to be looked at separately, the authors working with the new setup and the readers of salto.bz. For the readers the new faceted search provides an obvious added benefit. They can now more easily search through the archive of news articles with the keywords helping them find all the articles on a specific subject.

For the authors, the added benefit is less immediate and only occurs if authors indeed can observe an effect of their extra work of assigning keyword pairs to their texts. By being tagged with keywords articles in the news archive are more likely to be found when readers search for related topics, and might in the future even be used to explicitly link older articles to the newly created ones. In this way, adding the keywords might increase the reach and longevity of an article. At the same time, authors have to (slightly) adapt their workflow by adding and checking the keywords which means additional work for them. Even though the amount is very tiny, authors, especially freelancers, already work under pretty tight deadlines and will try to avoid any delays, even if it is just waiting a couple of seconds for an external service to produce some keywords.

When evaluating the use case with our partners at salto.bz it became clear that already during the official run-time of the project the use of the keyword generator had slowly but steadily decreased.

5.4 Sustainability of the Cooperation

Given that the cooperation started as a project effort of limited duration and with limited resources the question of sustainability is crucial and has to always be kept in mind during the run-time of the project. The project outcome has to provide enough added value for both sides to continue to invest time to keep it running smoothly, while at the same time the use case should be set up in such a way that, once everything is implemented, the amount of maintenance to keep it running is minimal.

As discussed above in 5.3, the added value is significant, but not equally distributed among all parties involved. While for the IAL and the readers of salto.bz the added value is immediate, the added value for the authors - increasing the longevity of their articles - can only be measured once the system has been running for some longer time. This leads to the unfortunate situation that the authors are the ones that have to continuously invest time into keyword validation, while being the ones for which the added value is least obvious at first.

Regarding the amount of maintenance needed for the system, the technical maintenance of the keyword generator should be fairly minimal, mostly consisting of keeping the server running and providing occasional security fixes if necessary. The same should be true for the Drupal UI modifications, though here any general Drupal updates might make also changes to the keyword UI necessary.

Apart from the technical maintenance, one larger maintenance task is the regular curation of keywords. To ensure the quality of the service in the long term, the news portal personnel has to regularly merge and check newly introduced keywords (see Section 5.2 above). We assume that the set of keywords will naturally consolidate over time to a certain extent, which means that the amount of work involved in merging and correcting will continuously decrease, but never reach zero.

A possible solution we have discussed to reduce this maintenance workload is to move to a semi-static model: A closed set of keywords is set at a given time, and only keywords from this closed set are assigned to new articles. This closed set is occasionally updated to include the most relevant new keywords ("hot topics") that occur over a more extended time.

As described above, the adoption of the system by the salto.bz authors decreased significantly during the project's run-time and we assume it would need new incentives to keep their engagement up. We therefore have decided to move the system into a more automatic state, where keywords from a closed list will be assigned automatically to new texts, removing the added work from the author while still

maintaining a system that assigns keywords to all new texts. In the mid term it is foreseen to revisit the implementation of the keyword extraction tool with the aim to provide an even more reliable service, which would naturally reduce the amount of manual curation needed.

6 Conclusions

The reported work on the use case helped better understand what is needed to establish infrastructure cooperations with non-research partners and what are particular challenges.

The most noteworthy challenge we encountered relates to establishing a cooperation and defining common use cases. We observed that besides a lack of awareness about ongoing language infrastructure initiatives, understanding what it can bring in terms of added value is missing on the side of non-research language actors. Resolving this issue requires extended interdisciplinary communication efforts to identify real needs of business partners and map them to existing LTI solutions. In order to start from tangible scenarios it would be desirable that infrastructure initiatives like CLARIN worked towards a portfolio of use cases for cooperations with non-research stakeholder groups. The recently started ENRIITC project¹⁶(McEntee, 2022) looks like it might be making steps in this direction, with its plan of establishing a network of Industrial Liaison and Contact Officers to facilitate cooperation between research and industry institutions. It also turned out to be highly relevant to reach a common understanding of the objectives and expectations in relation to the use case, as well as a detailed analysis of the workflow and roles needed to implement it right from the start of the project in order to create a sustainable initiative. In the specific use-case of integrating a keyword extraction service into a news portal we ended up with a workflow that was driven by several layers of indirection. While design decisions were taken by the project coordination and technical staff, the immediate contribution to make the keyword assignment happen was required from the news authors, and the results would benefit mainly the news readers. The missing alignment of expectations of all involved parties resulted in a reduced commitment from the side of the authors and thus undermined the sustainability of the effort (see 5.3 and 5.4 for some more details).

A second difficulty relates to integrating the technical implementation with established workflows of a partner institution and fostering the adoption of new procedures. Overall, the communication and decision-making processes required interactions well beyond the technical level and concerned management and editorial participation to a considerable extent. Also, because the realisation of the use case impacted the customer-facing newspaper portal search interface, many people outside our direct contacts paid great attention to all the changes. This experience shows that the workflow planning is a highly complex problem in itself even before detailed aspects of technical solutions or performance of its individual components come in. Again this shows that a highly interactive communication effort between all partners is inevitable and an asynchronous project cooperation with the major workforce on the side of the research partner is not realistically feasible. For future endeavours this suggests that truly interdisciplinary and inter-institutional project cooperations should be targeted, as mentioned above.

Finally, to create a sustainable service addressing questions of quality control and long term maintenance of the service become crucial. The fact that through this project, the news portal created a dependency on an external service as part of their daily workflows underlines the importance of sustainability of infrastructure services and strategies for long term maintenance, which both pose unresolved challenges, not only in this kind of interdisciplinary cooperation but in the whole field of technical infrastructure.

We conclude that up until today, establishing any infrastructure cooperations with non-research partners requires substantial efforts in communication, workflow planning and technical solution building on both ends, the research partner and the industry partner. As long as no portfolio of use cases and applications exist that can be re-purposed, successful cooperations can likely only be created in the context of bigger funded projects. They would need to bring together research and industry partners for several years of intense collaboration, since as of today, on the side of the non-research client of European infrastructures, both awareness for what is doable and support for implementation is still greatly lacking, while on the research side, technical solutions are often not at the level of being ready for the market without considerable customization.

¹⁶<https://enriitc.eu/>

Acknowledgements

We would like to thank Monica Pretti for the detailed manual analysis and correction of the first comprehensive sample of automatically generated keyword pairs and the harmonisation of newly added keyword pairs during a limited testing period.

References

- Bleichner, M., Giesbrecht, E., Gust, H., Leicht, E.-M., Ludewig, P., Möller, S., Müller, W., Schmidt, M., Stefaner, M., Stemle, E., and Wilke, K. 2005. *ASADO: The Analysis and Structuring of Aviation Documents - Final Report*, Institute of Cognitive Science at the University of Osnabrück and Institute of Applied Linguistics at the University of Hildesheim. <https://api.zotero.org/users/332053/publications/items/KSJ9ECLV/file/view>.
- de Jong, F. M. G., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, Dieter 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, European Language Resources Association (ELRA). <http://dspace.library.uu.nl/handle/1874/364776>.
- Edmond, J., Fischer, F., Mertens, M., and Romary, L. 2017. The DARIAH ERIC: Redefining research infrastructure for the arts and humanities in the digital age, *ERICIM News*, 111.
- Fielding, R. T. 2000. REST: Architectural styles and the design of network-based software architectures, *PhD thesis*, University of California, Irvine <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>.
- Fišer, D., Lenardič, J., and Erjavec, T. 2018. CLARIN's Key Resource Families, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, European Language Resources Association (ELRA). <https://aclanthology.org/L18-1210>.
- Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Ďurčo, M., and Schonefeld, O. 2014. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure, *Selected Papers from the CLARIN 2014 Conference*, p. 36–53.
- Lyding, V., König, A., Gorgaini, E., Nicolas, L., and Pretti, M. 2019. DI-ÖSS - Building a digital infrastructure in South Tyrol, *Selected papers from the CLARIN Annual Conference 2018*, Pisa, 8-10 October 2018 / edited by Inguna Skadina, Maria Eskevich, Linköping Electronic Conference Proceedings, 159(10), p. 92–102.
- Lyding, V., König, A., and Pretti, M. 2020. Digital Language Infrastructures—Documenting Language Actors, *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 3457–3462.
- McEntee, J. 2022. Building bridges between big science and industry, *Physics World*, 35(1), IOP Publishing, p. 8–9i, <https://doi.org/10.1088/2058-7058/35/01/10>.
- Poesio, M. and Magnini, B. 2009. Content Extraction Meets the Social Web in the LiveMemories Project, *Proceedings of the Workshop on Advanced Technologies for Digital Libraries 2009 (AT4DL 2009)*, Bozen Bolzano University Press, p. 42–45.
- Rehm, G., Piperidis, S., Bontcheva, K., Hajic, J., Arranz, V., Vasiljevs, A., Backfried, G., Gómez-Pérez, J., Germann, U., Calizzano, R., and others 2021. European Language Grid: A Joint Platform for the European Language Technology Community, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, p. 221–230.
- Woldrich, A., Goli, T., Kosem, I., Matuška, O., and Wissik, T., 2021. ELEXIS: Technical and social infrastructure for lexicography, *K Lexical News*, (28), Zenodo, p. 45–52, <https://doi.org/10.5281/zenodo.4607957>

Annotation Management Tool: A Requirement for Corpus Construction

Yousuf Ali Mohammed, Arild Matsson, Elena Volodina

University of Gothenburg, Sweden

name.surname1.surname2@svenska.gu.se

Abstract

We present an annotation management tool, `SweLL portal`, that has been developed for the purposes of the `SweLL` infrastructure project for building a learner corpus of Swedish (Volodina et al., 2019). The `SweLL portal` has been used for supervised access to the database, data versioning, import and export of data and metadata, statistical overview, administration of annotation tasks, monitoring of annotation tasks and reliability controls. The development of the portal was driven by visions of longitudinal sustainable data storage and was partially shaped by situational needs reported by portal users, including project managers, researchers, and annotators.

1 Introduction

During 2017–2021, we were setting up the foundation for empirically based research on Swedish as a second language. The results were released in 2021 under the name of `SweLL` infrastructure, as a part of Nationella Språkbanken and Swedish CLARIN.¹ The core work entailed collecting and manually annotating learner written essays, the `SweLL-gold` corpus (Volodina et al., 2019). However, this process turned out to be more complex and involved a lot of work “behind the scenes”. *First*, to make sure the annotations are reliable, we invested extensive work into developing and documenting a taxonomy of corrections (or errors, a more traditional term used in other projects) (Rudebeck and Sundberg, 2021) and a taxonomy of personally identifiable information for successful pseudonymisation (Megyesi et al., 2021). *Second*, to ensure the consistency of the manual annotation, we developed a tool to support the annotation itself, namely the `Svala` annotation tool (Wirén et al., 2019) and tool for the management of the annotation process, the `SweLL portal`. *Third*, to make sure the resulting collection of essays can reach the intended user, we worked on the legal aspects of access to the material as well as on the visualisation of the corpus so that it may be browsed and analyzed statistically based on textual, educational and linguistic characteristics.

From the above follows that an infrastructure project dealing with the construction and annotation of an electronic learner corpus entails the collection of data and metadata, followed by the meticulous selection of essays for manual annotation to ensure the balance and representativity of various metadata (e.g. the balance between texts of different genres and topics, between the writer’s gender and education level, etc.) and the annotation itself. There are four pillars that tend to be named in connection to digital infrastructures: data, tools for data annotation, tools for data exploration, and expertise (Volodina et al., 2016).² What is usually overlooked is some project management environment.

Fort (2016) and (Hovy and Lavid, 2010) emphasise the need for an annotation management software that would ensure the reliability of manual annotations. There are two main reasons for that: *First*, a corpus of good quality must boast representativeness of the language it embodies and balance of the samples that characterise the language. This requires monitoring the collected text instances with regards to the various types of metadata. *Second*, the data as such is only the first step, the most interesting

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

¹<https://spraakbanken.gu.se/en/projects/swell>

²<https://spraakbanken.gu.se/projekt/swell>

research can be done when the data is annotated for one or another text- or language-related feature, and this annotation should be reliable. 'Tools decay, data stay' is only true when data is reliably annotated.

The debate on the quality of data annotation often goes in the direction of (1) *tag sets* – their size and ambiguity, (2) *guidelines* – their clarity and degree of detail, and (3) *tools used for annotation* – their user-friendliness and support in annotation. The annotation management as such – database handling, statistical overviews, inter-annotator agreement controls, etc. – is often overlooked or simply not considered in time. This, then, results in an annotation project being managed using Excel files, which leads to errors, imbalance, loss of annotation or information and ultimately to the reduction of annotation quality (Stemle et al., 2019).

A number of data management tools have been developed in different projects. Most of them, however, were initially developed to support manual annotation, but with time added some functionality for database communication and versioning. Some examples of those are TEITOK (Janssen, 2016), WebAnno (de Castilho et al., 2014), the UAM Corpus Tool (O'Donnell, 2008). Such tools combine data annotation with data management, which makes them task-oriented and reduces their flexibility in the choice of an annotation paradigm. For example, in the case of TEITOK, the annotation is performed using xml TEI format, which may or may not be an optimal format, even though the data management functionality might satisfy a new project. Creating a universal tool that would satisfy any project (i.e. 'one size fits all') is no simple task. Due to the outlined considerations, we opted to develop our own tools, separating data management from data annotation.

The *SweLL-gold corpus* that we have been constructing over the past several years is aimed at researchers, developers and teachers to promote the fields of Second Language Acquisition (SLA), Language Assessment (LA), Intelligent Computer-Assisted Language Learning (ICALL) and Language Technology approaches to those – predominantly within CLARIN and other European (due to the GDPR restrictions) user groups. Due to this, a high standard of annotation is required. The *SweLL portal* is one of the steps to ensure those standards. Looking back at our experiences and analyzing the benefits of an annotation management tool, we can say that its use has helped in more ways than just corpus preparation. Among others, we have tested uploading other (bonus) learner corpora to the portal, and exporting them from the portal applying a unified set of metadata attributes and values (using 'N/A' as a value for absent attributes). This step has helped us make several Swedish learner corpora interoperable with each other, *interoperability* being a known challenge in CLARIN-related context (König et al., 2021; Stemle et al., 2019; Volodina et al., 2018).

The *SweLL portal* is deployed on the university servers at Språkbanken Text, Sweden, and only permits the storage of pseudonymised text data according to GDPR regulations. Due to this fact, we are restrictive about allowing free access to the *SweLL portal* for other users who may incidentally upload personally identifiable text data. Only approved users are added to the portal. The code for the portal is available at a GitHub repository³ for users in need of a data management tool.

Below, we describe the architecture of the *SweLL portal*, from data management to data import and export, and outline some current developments and future plans.

2 Data Management

The *SweLL portal* is a user-friendly tool for metadata and data collection and for annotation management. The three modules (*datacollection*, *task_manager*, *annotation*) in Figure 1 are loosely dependent on one another, so that another hypothetical project might replace only the *datacollection* and/or *annotation* modules with their custom implementations.

The *datacollection* module contains the *SweLL* metadata model. It is an interface that communicates with the database, where one can store, access and manage *metadata* about learners, tasks, schools and individual texts. New metadata records can be created following a standard form and stored in the metadata container. The metadata container consists of four objects:

1. *Source* stores information about the school where the essays have been collected. A *Source* is represented by a school ID, the type of education and the course type.

³<https://github.com/spraakbanken/swell-portal>

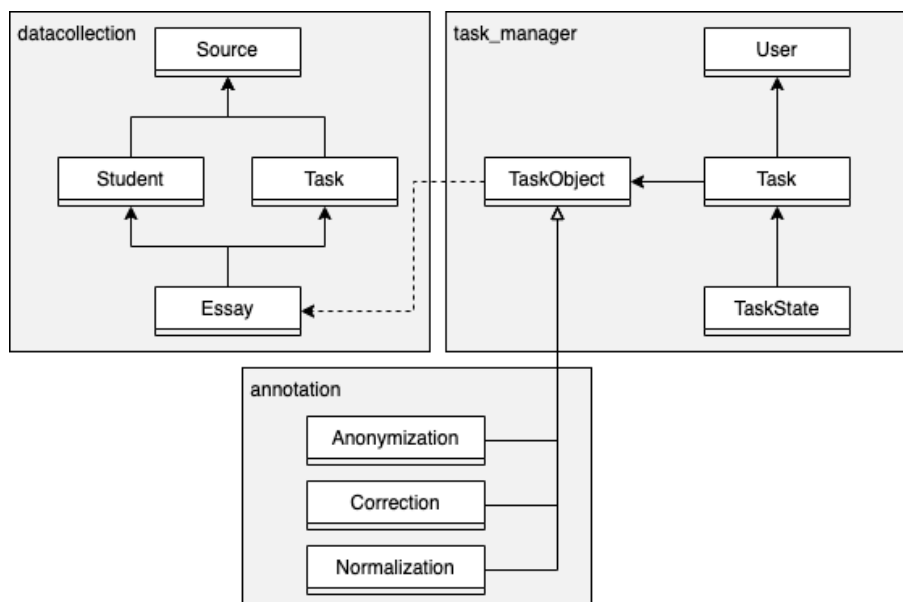


Figure 1: A partial class diagram of the application data model.

2. *Student* stores information about students, including student IDs, and their structured socio-demographic information about gender, mother tongue(s), education, etc.

3. *Task* stores information about the task learners have received for essay writing, including descriptive information about the genre, topic, grading system, allowed time, etc.

4. *Essay* metadata is a record created as a response by a *Student* to a *Task*, which also stores information about the individual performance on an essay. (Essay texts are not stored here, but in *TaskObject* and *TaskState* objects, see below.)

Skrivuppgifter

Filtrera Lägg till ny

ID	Titel	Datum	Nivå	Skolform	Genre/texttyp	Uppgiftstyp	Kurs	Betygsskala	Metastatus
AT2	Om din bostad och om att bo	2018-W08	Nybörjare	Vuxenutbildningen	Argumenterande	Inplaceringsprov	Inplaceringsprov SFI	SFI-Inplacering	Edit
AT3	Berätta hur du bor!	2018-W17	Nybörjare	Vuxenutbildningen	Beskrivande	Inplaceringsprov	Inplaceringsprov SFI	SFI-Inplacering	Edit
AT4	Om din bostad och om att bo	2018-W17	Nybörjare	Vuxenutbildningen	Argumenterande	Inplaceringsprov	Inplaceringsprov SFI	SFI-Inplacering	Edit
BT1	Utredande text (pm), övning inför NP	2018-W16	Avancerad	Ungdomsgymnasiet	Utredande	Formativ skrivuppgift	SVA 3	Uppgiften har inte betyg	Edit

Figure 2: A list of *Task* metadata records.

In the user interface, under the *Metadata* tab, it is possible to get an overview of all items in each of the four objects mentioned above (e.g. Figure 2⁴). For each object, there exists an option to filter the metadata based on different attributes, to open existing records for editing and to add new records (e.g. Figure 3).

3 Annotation Task Management

The annotation task management is based on two modules in Figure 1, namely the *annotation* module and the *task manager* module.

⁴Text in the Figures is predominantly in Swedish.

Personlig information

Swell-id

Kön Annat Kvinna Man Vill inte säga

Född (år) 2000–2004 1995–1999 1990–1994 1985–1989 1980–1984 1975–1979 1970–1974
 1965–1969 1960–1964 1955–1959 1950–1954 1945–1949 1940–1944 tidigare

Total tid i Sverige år månader

Utbildning

	Utanför Sverige		I Sverige	
Grundskola	<input type="text" value=""/> år	<input type="text" value=""/> månader	<input type="text" value=""/> år	<input type="text" value=""/> månader
Introduktionsprogram			<input type="text" value=""/> år	<input type="text" value=""/> månader
Gymnasiet	<input type="text" value=""/> år	<input type="text" value=""/> månader	<input type="text" value=""/> år	<input type="text" value=""/> månader

Figure 3: A record for *Student* personal metadata.

The *annotation* module consists of three tasks: Anonymisation,⁵ Normalisation and Correction Annotation. Each task type is defined through the workflow control that opens them in an external annotation tool, SVALA (Wirén et al., 2019), which is a stand-alone application. The SVALA code for all three tasks can be adapted to annotation of other languages than Swedish.

The *task_manager* module allows superusers to create, assign and manage annotation tasks using the *TaskObject* shown in Figure 1. A *TaskObject* pairs a task type with a specific essay, e.g. *anonymisation of essay AIAT2*. It is implemented with a generic reference to the essay metadata object, ensuring a loose module dependency. The *TaskObject* is a collection of three objects, *User*, *Task* and *TaskState*.

1. *User* is a record associated with an annotation expert performing the task. The record consists of an ID of a portal user who has been assigned the annotation task.

2. *Task* represents which annotation task is being performed (anonymisation, normalisation, correction annotation) and tracks a specific user’s work on a *TaskObject*. If more than one user work on the same annotation task, they each have a separate *Task* with separate progress.

3. *TaskState* shows three states of work on a *Task*: assigned, started and completed, each of which can have a Boolean *yes/no* flag. Work in the annotation tool generates a sequence of *TaskStates*, each a snapshot version of the text plus annotations.

When a *User* starts an annotation *Task*, the essay opens in the external annotation tool SVALA (Wirén et al., 2019)⁶. A unique version of the essay is saved in the *TaskState* on every introduced change by the annotator in the SVALA tool. Once the annotation task is completed, the task can be marked as *Done*. This updates the status of the *TaskState* in the SwELL portal.

The functionality of the portal allows the superuser to assign the same Correction Annotation task to several users. When two or more versions of the same correction annotation task are completed, it is possible to measure Inter-Annotator Agreement (Figure 4). There is a possibility to click on the EssayID on the Annotations page to view the full text and to monitor the progress of the annotation.

⁵Later we switched from using the term *Anonymisation* to use the term *Pseudonymisation*.

⁶SVALA demo version: <https://spraakbanken.gu.se/swell/dev/>

Inter-annotator agreement		Label stats			
Annotators	Annotator 1 Annotator 2	Annotator	Annotator 1	Annotator 2	Total
Krippendorff α	0,973	C	1	2	3
Average observed agreement	0,983	L-Der	1	1	2
Multi kappa (Davies & Fleiss 1982)	0,975	L-FL	1	1	2
		L-W	3	3	6
		M-Def	2	2	4
		M-Gend	1	1	2
		M-Num	1	1	2
		M-Verb	1	1	2
		O	18	18	36
		S-R	1	1	2
		S-Type	2	2	4
		S-WO	1	1	2

Figure 4: Inter-annotator agreement for a particular essay and a pair of annotators.

4 Annotation Tool and its Demo Version

The SVALA tool (Wirén et al., 2019) is a stand-alone annotation tool that was developed with the aim of supporting the manual annotation work on learner essays in a user-friendly way. Annotators can use this tool for different annotation tasks such as pseudonymisation, normalisation and correction annotation. The essays are shown in the original and target versions, and as a 'spaghetti' version. The spaghetti format maps the source text (*written by learners*) to the target text (*normalised by the annotator*) token by token and allows the annotators to add the labels (*pseudonymisation or correction*) to each edge in the graph. This tool also comes with an automatic pseudonymisation pipeline that can de-identify the personal information in an essay using rule-based methods.

SVALA demo version⁷ is a copy of the SVALA tool (Wirén et al., 2019) that is publically available for anyone interested in testing the annotation of their datasets. This version is not connected to the *SweLL* portal or any database, and is used for demo-purposes and for viewing the full content in the essays through the Korp corpus search tool (Ahlberg et al., 2013). Full texts that are opened in the demo version of SVALA can be modified without any risk of sabotaging the annotations on the server, since these changes are not saved to the database.

5 Statistics

The statistics section shows an overview of the metadata and its frequencies, as well as frequencies over tokens and sentences in the *SweLL-gold corpus*. Described below are two ways of viewing the statistics:

1. *Statistics*: On this page, one can view the statistics of the pseudonymisation and correction labels together with the number of tokens, correct sentences and incorrect ones in the *SweLL-gold corpus* as shown in Figure 5. The statistics for attributes in student, task and essay metadata are also given on this page. There is an option to download the statistics as a CSV file.

2. *Summary*: Metadata for students, tasks and essays as well as annotation data can be filtered and viewed in a table format as shown in Figure 6. To have a better understanding of the data one can even view the tables summarised into graphs. The statistics can be downloaded for further work either in CSV, plain text or JSON formats. The summary page also has an option to visualise and monitor the progress of the annotation work for the project.

A decision has been made in the project not to implement advanced search and filters. Instead, a possibility is provided to download files with statistics that can be opened using Excel or processed through other programs that offer advanced filtering.

⁷SVALA demo version: <https://spraakbanken.gu.se/swell/dev/>

	Average observed agreement	0,962
Korrigeringsannoterade (statistik)	Antal meningar	8481
	Antal meningar/upsats	17
	Antal meningar med fel	6657
	Antal korrigeringar/upsatser	53
	Antal korrigeringar/mening	3
Antal tokens (μ - mean, σ - standard deviation)	Anonymiserade uppsatser, källtext	211744 ($\mu = 317.0$, $\sigma = 241.4$)
	Normaliserade uppsatser, källtext	148905 ($\mu = 296.6$, $\sigma = 246.8$)
	Normaliserade uppsatser, målttext	152518 ($\mu = 303.8$, $\sigma = 250.9$)
	Korrigeringsannoterade uppsatser, målttext	152518 ($\mu = 303.8$, $\sigma = 250.9$)

Figure 5: Running statistics over all material in the portal (excerpt).

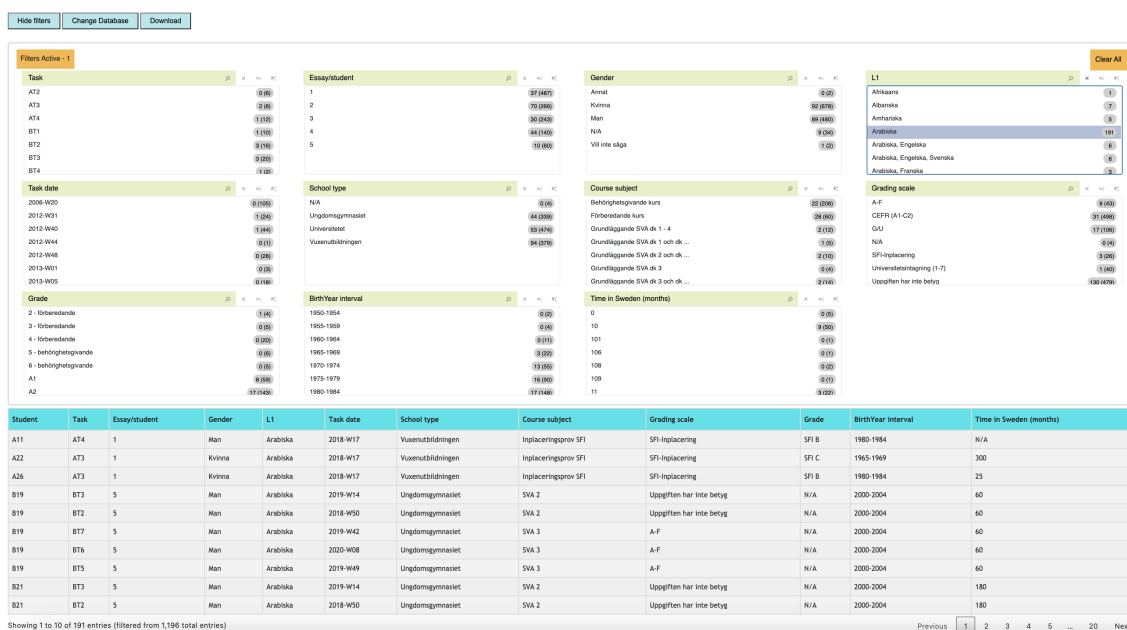


Figure 6: Statistics with a filter function.

6 Data Import and Export

In SweLL portal, there are two different methods to import the essays into the database, namely, an XML-import and a raw text import. The XML import has an additional functionality to automatically create and insert the metadata for a *Student*, *Task*, and *Essay* on importing an XML file. The metadata attributes are part of the header tag in the XML file. In raw text import, the metadata for *Student*, *Task*, and *Essay* should be created manually in advance to correctly store the raw text. On completion of the import, successfully imported essays are ready for annotation in the task manager module.

The data export functionality gives users the ability to export the data based on a variety of selections as shown below.

- Annotation type - normalisation or correction annotation
- School - school ID (A,B,C...)
- Mother tongue - Arabic, English etc.
- Status - complete, incomplete or both

- Data type - source or target
- File type - XML, JSON or plain text.

There is a script for converting the working format based on JSON into XML that can be used for import of the data into the Korp search tool (Ahlberg et al., 2013).

Access to the SweLL portal is password-protected and only users with proper access rights can import and export the data. The work on documenting the SweLL portal is ongoing, with some documentation available from the SweLL project webpage.⁸

7 Future

The SweLL portal at the moment of release contained 502 essays with correction annotation and approx. 700 essays without correction annotation. In the past six months, the collection has grown with approximately 50 more essays which are in the process of manual normalisation and correction annotation. We expect it to grow further since we are currently collaborating with two international teams who are reusing our tools: a Slovenian team and a French team. Apart from enriching our data, the collaboration brings into spotlight aspects that we had not previously considered, including the expansion of metadata types, e.g. to cover individual differences (modern aptitude tests), more refined taxonomy for tasks (e.g. what linguistic parameters they stimulate, as a support for language assessment and cross-comparison of tasks), new taxonomy for content-based feedback and feedback on linguistic features/errors; etc.

We are also planning to provide a possibility for teachers/researchers to visualise the progress of a subcorpus (e.g. a Class) based on error correction labels. This will help them to identify the necessary focus for the learners, or to see whether or not the learners are making progress from one written task to the next.

Our further plans include:

1. Making the data statistics available for non-login users
2. Adding new types of users, e.g. teachers and classes
3. Creating subfolders for the data inside a project so different groups can work and collaborate
4. Creating a possibility for different projects to work independently from each other
5. Visualising learner/groups/class progress over time.

8 Concluding Remarks

In the current project, the aspects of data storage and annotation management have been taken seriously following the arguments outlined in (Fort, 2016) and (Hovy and Lavid, 2010) that stable annotation management is one of the important prerequisites for the creation of well-balanced and reliably annotated corpora. The portal development was incremental, with changes introduced in response to the needs of the project. Overall, both project researchers and project assistants were aided in their work through the SweLL portal functionalities. The SweLL portal will continue to be used for new learner corpus annotation projects as well as for statistical exploration of the material as a part of a newly developed SweLL infrastructure for second language research.

Given the architecture of the main components in the portal (Figure 1), some further level of abstraction can be added to it, so that the approach and the framework could be applied to a broader scope of corpus annotation projects. That would entail, for example, scenarios such as an abstracting database/*datacollection* module, a module for *annotation* types/states and a *task_manager* on the one hand; and adding flexibility for the integration of an external annotation tool, on the other; with an active visualisation module and statistics tool.

⁸<https://spraakbanken.gu.se/en/projects/swell/swell-docs>

Acknowledgements

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *SweLL - research infrastructure for Swedish as a second language*, IN16-0464:1, and by *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We also wish to thank the anonymous reviewers for their valuable comments on a previous version.

References

- Ahlberg, M., Borin, L., Forsberg, M., Hammarstedt, M., Olsson, L.-J., Olsson, O., Roxendal, J., and Uppström, J. 2013. *Korp and Karp - a bestiary of language resources: the research infrastructure of Språkbanken*, 429–433. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013).
- de Castilho, R. E., Biemann, C., Gurevych, I., and Yimam, S. M. 2014. *WebAnno: a flexible, web-based annotation tool for CLARIN*. Proceedings of the CLARIN Annual Conference (CAC). Citeseer.
- Fort, K. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.
- Hovy, E. and Lavid, J. 2010. *Towards a ‘science’ of corpus annotation: a new methodological challenge for corpus linguistics*, 22(1):13–36. International journal of translation.
- Janssen, M. 2016. *TEITOK: Text-faithful annotated corpora*, 4037–4043. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16).
- König, A., Frey, J.-C., and Stemle, E. W. 2021. *Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora*, 12(5):199. Information. Multidisciplinary Digital Publishing Institute.
- Megyesi, B., Rudebeck, L., and Volodina, E. 2021. *SweLL pseudonymization guidelines*. Technical report. GU-ISS Forskningsrapporter från Institutionen för svenska språket (2011-), University of Gothenburg.
- O’Donnell, M. 2008. *Demonstration of the UAM CorpusTool for text and image annotation*, 13–16. Proceedings of the ACL-08: HLT Demo Session.
- Rudebeck, L. and Sundberg, G. 2021. *SweLL correction annotation guidelines*. Technical report. GU-ISS Forskningsrapporter från Institutionen för svenska språket (2011-), University of Gothenburg.
- Stemle, E. W., Boyd, A., Jansen, M., Lindström Tiedemann, T., Mikelić Preradović, N., Rosen, A., Rosén, D., Volodina, E., et al. 2019. *Working together towards an ideal infrastructure for language learner corpora*. Presses universitaires de Louvain.
- Volodina, E., Megyesi, B., Wirén, M., Granstedt, L., Prentice, J., Reichenberg, M., and Sundberg, G. 2016. *A Friend in Need? Research agenda for electronic Second Language infrastructure*. Proceedings of Swedish Language Technology Conference (SLTC) 2016, Umeå, Sweden.
- Volodina, E., Jansen, M., Stemle, E. W., Lindström Tiedemann, T., Mikelić Preradović, N., Ragnhildstveit, S. K., Tenfjord, K., and Koenraad, D. 2018. *Interoperability of Second Language Resources and Tools*, 90–94. Proceedings of the CLARIN Annual Conference 2018, Pisa, Italy.
- Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice, J., Rosén, D., Rudebeck, L., Schenström, C.-J., Sundberg, G., and Wirén, M. 2019. *The SweLL Language Learner Corpus: From Design to Annotation*. Northern European Journal of Language Technology, Special Issue.
- Wirén, M., Matsson, A., Rosén, D., and Volodina, E. 2019. *SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora*. Linköping University Electronic Press. Proceedings of CLARIN 2018.

Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS

Anna Björk Nikulásdóttir¹, Þórunn Arnardóttir², Starkaður Barkarson³,
Jón Guðnason⁴, Þorsteinn Daði Gunnarsson⁴, Anton Karl Ingason²,
Haukur Páll Jónsson⁵, Hrafn Loftsson⁴, Hulda Óladóttir⁵, Eiríkur Rögnvaldsson²,
Einar Freyr Sigurðsson³, Atli Þór Sigurgeirsson⁶, Vésteinn Snæbjarnarson⁵,
Steinþór Steingrímsson³, Gunnar Thor Örnólfsson⁴

¹Grammatek ehf., Iceland, ²University of Iceland, ³The Árni Magnússon Institute for Icelandic Studies,
⁴Reykjavik University, ⁵Miðeind ehf., Iceland, ⁶University of Edinburgh
anna@grammatek.com, thar@hi.is, starkadur.barkarson@arnastofnun.is
jg@ru.is, thorsteinng@ru.is, antoni@hi.is, haukurpj@miðeind.is,
hrafn@ru.is, hulda@miðeind.is, eirikur@hi.is,
einar.freyr.sigurdsson@arnastofnun.is,
atlisigurgeirsson@gmail.com, vesteinn@miðeind.is,
steinthor.steingrimsson@arnastofnun.is, gunnaro@ru.is

Abstract

In this paper we describe how a fairly new CLARIN member is building a broad collection of national language resources for use in language technology (LT). As a CLARIN C-centre, CLARIN-IS is hosting metadata for various text and speech corpora, lexical resources, software packages and models. The providers of the resources are universities, institutions and private companies working on a national LT infrastructure initiative, Language Technology Programme for Icelandic. All deliverables of the programme are published under open licences and are freely accessible for research as well as commercial use. We provide a broad overview of the available repositories and the core publishing guidelines.

1 Introduction

With the enormous progress in language technology (LT) in the last decades, the use of LT in research and commercial products has greatly increased. LT tools and resources are now not only used by LT specialists but also by researchers and developers from various fields. Beside the improvement in quality and usability, this development is driven by open access to data and software. For such resources to be of broad use, they need to be easily accessible and thoroughly documented. Thus, the large national LT infrastructure initiative *Language Technology Programme for Icelandic (LTPI) 2019–2023* (Nikulásdóttir et al., 2020b) chose CLARIN-IS to be the central hub for all deliverables of the programme.

This paper gives a broad overview of the manifold “buffet“ of available repositories and the core publishing guidelines.

2 CLARIN-IS

Iceland became a CLARIN ERIC member on February 1, 2020 after having an observer status since November 1, 2018. The Árni Magnússon Institute for Icelandic Studies is the leading partner in the Icelandic national consortium. The main motivation for joining CLARIN was to have a secure and well recognized infrastructure to store all the resources and tools created during the LTPI. But the plan for

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

the near future is to widen the scope and reach out to researchers of humanities and social sciences. A Metadata Providing Centre (CLARIN C-centre¹) has been established at the institute that hosts metadata for Icelandic language resources and distributes them through a Virtual Language Observatory.

As a new member, CLARIN-IS is in the process of establishing a technical Service Providing Centre (CLARIN B-centre), which will maintain language resources among other tasks. For now, we maintain a Gitlab², where all relevant GitHub repositories are mirrored, and deliver all resources to the C-centre.

3 Language Technology Programme for Icelandic

In October 2019, a consortium of Icelandic universities, companies and institutions (10 in total) started working on the LTPI. The programme aims at making Icelandic viable in future technologies that rely on LT in one way or another. To build foundations for that goal, the LTPI concentrates on developing language resources and infrastructure software, divided into six core project areas:

1. Language Resources
2. Support Tools
3. Machine Translation
4. Spell and Grammar Checking
5. Automatic Speech Recognition
6. Speech Synthesis

Each project area was further divided into work packages with defined goals. In total, 65 work packages were described up front with estimated 1136 man-months over five years to deliver the projects. The work packages have been revised and defined in more detail every year to keep up with developments in the field and to adjust work packages within project areas with the aim of meeting the overall goals of the programme. During the preparation work on the LTPI, other European national programmes for LT were reviewed and information from experienced partners collected. Further information on related programmes and the general structure and execution of the LTPI can be found in (Nikulásdóttir et al., 2020b).

All deliverables of the programme are published under open licences and are freely accessible for research as well as commercial use. Therefore, it is of utmost importance to have a stable hosting platform that can ensure access and availability.

In the remainder of this section, we describe each of the core projects, along with evaluation results where applicable.

3.1 Language Resources

A variety of language resources are being compiled or extended within the LTPI. The following list describes the main resources.

- The **Icelandic Gigaword Corpus (IGC)**³ is a large text corpus containing texts from various sources: news media, parliamentary proceedings, published books, journals, adjudications and more. The first version, published in 2018, contained over 1.2B running words from texts published until the end of 2017 (Steingrímsson et al., 2018), while the latest version contains close to 1.9B words (Barkarson et al., 2022). Within the LTPI, the corpus is being updated yearly with new data sources and updated data from previous ones. Each new edition is annotated using the latest tools. Table 1 shows the development of the corpus, year by year.

¹<https://clarin.is/>

²<https://gitlab.com/icelandic-lt>

³<http://igc.arnastofnun.is/>

Version	Words (M)	PoS tagger	Tagset
IGC-2018	1,253	IceStagger (Loftsson and Östling, 2013)	MIM-GOLD 1.0
IGC-2019	1,394	IceStagger	MIM-GOLD 1.0
IGC-2020	1,555	ABLTagger 0.9 (Steingrímsson et al., 2019)	MIM-GOLD 2.0
IGC-2021	1,871	ABLTagger 2.0 (Jónsson et al., 2021)	MIM-GOLD 2.0

Table 1. The table shows the amount of tokens (millions) for the four published versions of the IGC, as well as the PoS taggers and tagsets used.

The corpus is published under two different licences. Approximately half of the corpus uses CC BY 4.0, while the other half is published under the MIM-licence, a custom licence developed for Icelandic text corpora to use in cases where the publishers of the texts cannot accept the terms of CC BY 4.0. Both licences allow use of the data for all research and language modelling.

The first three versions of the corpus were published in two parts, one for each licence. As of version IGC-2021, the corpus is split into eight subcorpora that reflect the different source types: journals, published books, parliamentary speeches, laws, adjudications, social media and two news corpora.

Evaluation sets have been released to evaluate the accuracy of PoS tagging of different text types (Barkarson et al., 2020). This can be used to evaluate the tagging accuracy of different subcorpora. Using ABLTagger 0.9 (see Section 3.2) the accuracy ranges from 94.34% to 97.79%, depending on text type.

- **MIM-GOLD** (Helgadóttir et al., 2014) is a corpus of one million tokens, manually annotated with PoS tags. Within the LTPI, manually checked lemmas have been added to the corpus and the tagset has been revised in order to be able to accommodate for URLs and symbols like emoticons that are common in some modern texts and to make clearer distinctions on how to tag proper nouns, foreign words, abbreviations and more, described in (Barkarson et al., 2021a). A version of this corpus, **MIM-GOLD-NER**, in which named entities (NEs) have been annotated, has also been made available. In MIM-GOLD-NER, about 48,000 Icelandic NEs are tagged with one of eight NE types (Ingólfssdóttir et al., 2020).
- The LT part of the **Database of Icelandic Morphology** (DIM), a multipurpose linguistic resource, has been further developed within the LTPI. The Database of Modern Icelandic Inflection (DMII), which has been in development since 2002, and comprises approx. 300,000 inflectional paradigms (Bjarnadóttir et al., 2019a), is accessible at CLARIN-IS (Bjarnadóttir, 2019), together with valency structures of verbs (Bjarnadóttir, 2021), a list of common abbreviations in Icelandic texts (Bjarnadóttir and Ingimundarson, 2021) and the DMII Core (Bjarnadóttir et al., 2019b), which contains the core vocabulary of contemporary Icelandic. The database of inflectional paradigms has been compressed and encapsulated in a Python package to facilitate quick lookup in programs (Þorsteinsson et al., 2021b).
- **Skiptir** (Rúnarsson, 2020) is a simple command line tool that uses Pyphen to hyphenate text. Along with the tool a new hyphenation dictionary was compiled (Rúnarsson et al., 2020).
- The new **Icelandic Word Web** (Dánielsson et al., 2021) is an LT-focused redesign of a database of semantically related entries. It is stored in a single RDF file accessible directly through CLARIN-IS (Jónsson et al., 2020c).
- Two evaluation sets for word embeddings have been adapted to Icelandic. **IceBATS** (Friðriksdóttir et al., 2021) is an Icelandic adaptation of the Bigger Analogy Test Set (BATS). It contains 98,000 analogy questions that cover inflectional and derivational morphology as well as lexicographic and

encyclopedic semantics (Dánielsson et al., 2022). An Icelandic version of **Multi-SimLex** (MSL) has been compiled. MSL is an evaluation protocol and associated dataset for lexical semantics (Dánielsson et al., 2021). The original English-language MSL builds on several older, well-known datasets, most notably SimLex-999, and has been released in more than a dozen languages.

3.2 Support Tools

Several NLP tools have been or are currently being developed or improved upon within the LTPI. Each tool is either used as part of a processing pipeline, or as a stand-alone tool. Here, we have focused on the tools that are the most helpful for more complex project areas currently within the LTPI, such as spell and grammar checking, in order to maximize the use of time and effort.

- **Tokenizer:** A tokenizer (Þorsteinsson et al., 2021d) has been developed that converts input text to streams of tokens, where each token is a separate word, punctuation sign, number/amount, date, e-mail, URL/URI, etc. It also segments the token stream into sentences, considering various corner cases of abbreviations, dates, etc. to prevent wrong segmentation. It reaches 100% accuracy for sentence detection for texts without (well-documented) edge cases, and 99.7% for token detection. Two modes of tokenization were implemented to serve the widest audience. PoS tagging and machine translation use the *shallow* tokenization, where tokens are separated by white space. Parsing and grammar correction rely on the *deep* tokenization, where the tokens have been annotated with the token type and further information extracted from the token.
- **PoS tagger:** Before the LTPI started, the best performing PoS tagger for Icelandic was ABLTagger 0.9, a BiLSTM model implemented in DyNet, achieving an accuracy of 94.47% when evaluated on the MIM-GOLD corpus with the original tagset (Steingrímsson et al., 2019). During the LTPI, this tagger has been gradually improved. First, it was ported to PyTorch and several parts of it improved, e.g. by adding pre-trained word embeddings (trained on the IGC), resulting in ABLTagger 1.0 on CLARIN-IS, which obtains an accuracy of 95.59% on the revised MIM-GOLD tagset, which all subsequent tagging models use. Second, by incorporating contextualized word embeddings, i.e. ELECTRA-Small trained on the IGC, resulting in ABLTagger 2.0 in CLARIN-IS (Jónsson et al., 2021), the accuracy has increased to 96.95%. Finally, by incorporating larger BERT-like models, e.g. ELECTRA-Base (Clark et al., 2020), the accuracy increases significantly, to 97.71%. This accuracy score refers to a model excluding the tags for non-analysed tokens (x) and foreign words (e).
- **Lemmatizer:** With resources from Section 3.1, MIM-GOLD and DIM, a RNN lemmatizer accepting the word form as well as the corresponding PoS tag to predict the lemma has been developed (Jónsson and Loftsson, 2021). Latest experiments show an accuracy of 98.9% on known lemmas and 86.6% on unknown lemmas.
- **Named Entity Recognizer:** In parallel to the construction of MIM-GOLD-NER (see Section 3.1), three different machine learning models were evaluated (Ingólfssdóttir et al., 2020). The best performing model was a bidirectional LSTM (BiLSTM) model, obtaining an overall F_1 -score of 83.9, for the entities `Person`, `Organization`, and `Location`. In the LTPI, we experimented with fine-tuning BERT-like models using MIM-GOLD-NER, as well as developing a combination method (Guðjónsson et al., 2021). By fine-tuning an ELECTRA-Base model, trained on the IGC, the F_1 -score increased dramatically to 91.9. An even higher F_1 -score was obtained by fine-tuning a RoBERTa-Base model, trained on the IGC and data from several other sources (Snæbjarnarson et al., 2022), i.e. 92.7. Combining three BERT-like models, using simple voting in CombiTagger (Henrich et al., 2009), further increased the F_1 -score to 93.2.
- **Parsers:** Two previously published **parsers** have been updated within the LTPI, a *full parser* and a *shallow parser*. The rule-based full-constituency parser (Þorsteinsson et al., 2021c) relies on a wide-coverage context-free grammar (CFG) and uses a parsing system based on an enhanced Earley parser (Þorsteinsson et al., 2019). The grammar contains over 5,600 nonterminals, 4,600 terminals and 19,000 productions in fully expanded form. It also gives feature agreement constraints for

gender, number, case, and person. An enhanced Earley-based parser generates ranked parse trees in shared packed parse forests. The parser is the foundation for the grammar checking module in the spell and grammar checker. The work in the second year led to the tool reaching an F-measure of 81.2.

The shallow parser, IceParser, is useful as a faster, lighter option for basic parsing, where a full parse is not required, for example in information extraction. The parser, which consists of a sequence of finite-state transducers, accepts PoS-tagged input and generates output according to a shallow syntactic annotation scheme (Loftsson and Rögnvaldsson, 2007). The work on the shallow parser consisted of making it accept tagged text according to the new MIM-GOLD tagset (see Section 3.1) and improving individual components. Evaluation shows that the new version IceParser 1.5.0 on CLARIN-IS (Loftsson et al., 2021) obtains an F-measure of 96.3 for phrases and 83.1 for syntactic functions.

- **Lexicon Acquisition Tool:** ALEXIA (Friðriksdóttir et al., 2021; Friðriksdóttir and Jasonarson, 2021) is used to find neologisms as well as other words that are more frequently used than before. It processes the IGC, but can also be adapted to other data sources. It returns a word list with relevant information, such as frequency per word form.

All the above tools are currently available through CLARIN-IS. By the end of the LTPI, we will also have added pre-trained embeddings, a Universal Dependencies (UD) parser, and BERT-like language models to CLARIN-IS, thus ensuring open access to the most important basic support tools for LT.

A few UD parsing models will be trained using GreynirCorpus (Þorsteinsson et al., 2021e), which is originally a constituency treebank but is being converted to the UD annotation scheme using UD-Converter. UDConverter is a tool that has already been used to create two Icelandic UD treebanks by converting constituency treebanks based on the Penn Treebank (Arnardóttir et al., 2020).

Other, more peripheral resources have been added to CLARIN-IS, such as a parsed corpus with a tree search program to search for specific syntactic structures (Þorsteinsson et al., 2021e), and a test suite for different parsing schemas using the parsed corpora.

3.3 Machine Translation

Machine translation (MT) is a substantial part of the LTPI. In its first year, we focused on assessing methods, gathering data and building up infrastructure. Several deliverables were developed as part of this effort and published on CLARIN-IS to share between parties of the consortium. In the second year, the best model methods were further improved and iterated on and collaboration with industry was explored. In addition, organizers of The Sixth Conference on Machine Translation (WMT21) (Barrault et al., 2021) were approached regarding adding Icelandic–English as one of the language pairs in the news translation competition. This was approved and the creation of the datasets used was funded by the LTPI.

Several **resources for MT** between Icelandic and English have been developed and released.

- **ParIce** (Barkarson and Steingrímsson, 2019) is a collection of parallel English–Icelandic corpora suitable for training MT systems. It contains texts from various sources, including the Bible, the European Medicines Agency, open-source software projects, OpenSubtitles, the Nordic Council of Ministers, the European Space Observatory and most substantially European Economy Area regulations (Steingrímsson and Barkarson, 2021). Development and test sets for five different subcorpora (Barkarson et al., 2021b) were labelled and manually reviewed.
- **English-Icelandic glossary** (Steingrímsson et al., 2021) contains over 230K English-Icelandic pairs, single words and multiword units, with probability scores for translations in both directions. The glossary was built using automatic methods for compiling candidate lists, which were then manually checked by human annotators or compared to available manually curated dictionaries and word lists.
- **IPAC** (Símonarson and Snæbjarnarson, 2021) is a parallel corpus extracted from student theses abstracts that cover a wide range of academic topics. This dataset is diverse in its subject matter and

contains text in somewhat complex language. It is suitable both for training and as a baseline test set for evaluating general translation performance.

- **Backtranslations** are synthetic parallel corpora created using existing translation systems that have been shown to be greatly beneficial when training neural translation models. Creating the translations requires substantial computational power, so the data has been released publicly (Símonarson et al., 2021a). The data was collected from Wikipedia, legal documents and news articles.
- **Synthetic corpora** with injected proper names were created and released. These contain parallel sentences with names (Símonarson et al., 2020) and entities (Jónsson et al., 2021) substituted and labelled. These are useful for injecting vocabulary and improving performance when translating proper names that should not be translated directly.

Three different MT methods were tried and tested in the first year to understand how more traditional methods and recent advances compared with the available data. The models are available on CLARIN-IS.

- (i) **Moses** is a non-neural statistical MT system which has been shown to perform well in low-resource settings with limited computing power.
- (ii) **BiLSTM** is a neural MT architecture which was among the best some years ago.
- (iii) **Transformers** are state-of-the-art neural MT models for widely used languages with high resources.

All models were compared and evaluated as described in (Jónsson et al., 2020b). The Transformer model showed best results indicating that MT for English-Icelandic is not limited by the amount of available data, i.e. it is possible to make use of the same state-of-the-art methods as for e.g. English-German translations.

In the second year we built on the experience of the first year and more recent developments in NMT. A multilingual language model, **mBART-25** (Tang et al., 2021), was used as a starting point and then fine-tuned for translation between Icelandic and English. This is described more thoroughly in (Símonarson et al., 2021b). The resulting models (Snæbjarnarson et al., 2021) are much improved translation models, and the backtranslation corpus was regenerated using this data. A command line interface has been made available for translation that fetches the necessary models from CLARIN-IS⁴. Finally, the best NMT system for translation between English and Icelandic has been adapted for translation of EEA regulations and has undergone testing at the Translation Center of the Ministry of Foreign Affairs in Iceland. Initial results are good, and the collaboration has been extended for another year.

Besides the datasets collected and models trained, some infrastructure development has taken place to support the translation projects. A **web-based translation interface** was created and set up online to compare the different models, along with translations provided by Google. This served as a way to compare translations between the participating organizations and allow for open discussion about evaluation. The code for the website⁵ was packaged and released on CLARIN-IS. **Model serving infrastructure** was implemented for the different methods, and code and configurations to deploy and run translations are distributed on CLARIN-IS (Snæbjarnarson et al., 2020; Jónsson et al., 2020a).

3.4 Spell and Grammar Checking

The work in this core project has focused on developing the necessary data and tools for detecting, categorizing and correcting errors for different user groups. Several resources are currently available through CLARIN-IS.

An annotated **general error corpus**, the Icelandic Error Corpus, uses a fine-grained error classification that facilitates performance measurements of the spell and grammar checking software (Arnardóttir et al., 2021). The error corpus consists of three text genres: student essays, online news text and Wikipedia

⁴The code is available in <https://github.com/mideind/GreynirSeq> and the package `greynirseq` is available in PyPI.

⁵See <https://velthyding.is>.

articles. These texts were previously published, without error annotation, as part of the Icelandic Gigaword Corpus. All texts are proofread and errors annotated according to the annotation scheme, which consists of three hierarchical levels: main categories, subcategories, and error codes. The error codes are used when annotating errors, but the main categories and subcategories are used when improving the spell and grammar checker. The corpus is split into a development and test set to enable its usage when developing the spell and grammar checker. The corpus consists of 4,044 texts with a total of 44,268 revisions and 56,794 unique errors. The average number of errors per 1,000 words in the corpus is 45.76, but this count varies depending on text genre.

Three **specialized error corpora**, each representing a particular user group, have been annotated and published in order to measure the software's performance on errors particular to the respective user groups. The corpora are created using the same methods as used when creating the general error corpus and the same annotation scheme is used in all cases. Texts included in the specialized error corpora are collected particularly for this purpose, so more information on the authors can be obtained with their consent, e.g. their age, native language and name, if they do not wish to be anonymous. The Icelandic L2 Error Corpus is a collection of texts written by second-language learners of Icelandic (Glišić and Ingason, 2021; Ingason et al., 2021c). The corpus consists of 76 texts in which 21,842 errors have been annotated. The authors of the texts are of 16 different nationalities, the most common ones being English and Filipino. The Icelandic Dyslexia Error Corpus is a collection of texts written by native Icelandic speakers with dyslexia (Ingason et al., 2021b). The corpus consists of 26 texts, wherein 5,730 errors have been annotated. The Icelandic Child Language Error Corpus (Ingason et al., 2021a) is the final corpus belonging to the specialized error corpora. It is a collection of texts written by native Icelandic speakers aged 10 to 15 and consists of 119 texts with 7,817 annotated errors. All texts in this corpus are published anonymously.

In addition to these error corpora, various **word lists and language models** were created to further improve the spell and grammar checker. They include aggregated error data from different sources, a database of confusion sets and a trigram language model to help with suggestions for corrections, and are the following:

- A list of Icelandic words that may in some way be considered inappropriate, taboo and/or loaded in use or meaning (Sólmundsdóttir et al., 2021). The list also includes words that are not very inappropriate but can be considered an unfortunate topic for children or questionable depending on context. The words are grouped together in categories depending on either their meaning, form or use.
- A list of common misspellings and their corrections was also created (Arnardóttir and Ingason, 2020a). The word forms originate from the development set of the general error corpus and were annotated as nonwords.
- Yet another word list, created for developing the spell and grammar checker, is a list of automatically prepared word forms containing systematic errors, along with their corrections (Arnardóttir and Ingason, 2020b). The list was prepared using a word list from DIM, which includes different Icelandic words and their inflections. In all cases, one item in the word form is changed, i.e. an accent removed from a letter or added to a letter, or a letter replaced by another letter. These particular errors are common misspellings made in Icelandic text.
- Three datasets related to errors in place names were also prepared. The datasets are in JSON format and encoded in UTF-8. The datasets are *isprep4isloc* (Þórðarson, 2020c), which contains the correct prepositions for various Icelandic place names, *isprep4cc* (Þórðarson, 2020b), which contains prepositions for various countries and autonomous territories, and *cities_is2en* (Þórðarson, 2020a), which maps city names in Icelandic to their English counterparts. By using the last dataset, city names in English can be translated to Icelandic when correcting text.
- A list of systematic inflectional errors was also created. It consists of common erroneous inflectional

rules, which have been applied to entries in DIM to produce the error entries. The list is stored as a config file within the checking software.

- In addition to these word lists, a pre-existing trigram language model, Icegrams (Þorsteinsson and Óladóttir, 2020), was re-trained, fine-tuned and expanded to include a wider selection of high-quality texts from the Icelandic Gigaword Corpus. The spell checker was used to correct the trigram data beforehand in a bootstrapping manner, as there were cases where very frequent errors were even more frequent than the correct version in the data. This ensured high-quality trigrams, and that the model did not suggest these errors at the expense of the correct version.

The **spell and grammar checking software** (Þorsteinsson et al., 2021a) is a Python package and command line tool. The version currently available on CLARIN-IS offers token-level correction and some grammar correction. The checker is available on the web⁶.

The token-level correction relies on the tokenizer from the support tools (see Section 3.2), along with the aforementioned word lists and language models. The basic tokenizer output, i.e. text split into sentences and tokens, is sent through an error tokenization layer. This layer detects context-independent token-level errors such as duplicated words, single words erroneously written as two or more, and phrases erroneously written as single words.

The sentence-level correction relies on the full-constituency CFG parser from the support tools. One of the first layers provides information on all possible tags and lemmas with the help of DIM and built-in compound analysis. After that, the checker can detect more complex token-level errors, such as capitalization errors and taboo words. Semi-fixed phrases and common erroneous variations are handled with a list of lemmatized forms of all words in the phrase. The trigram model is used to find all possible substitutes for unknown or rare words, ranked by likelihood. All error tags and possible corrections are attached to the corresponding tokens.

The parser chooses the tag that best fits the context and results in a valid syntactic structure. In order to handle known, invalid structures in the grammar checker, special rules were added to the context-free grammar to capture those structures and in some cases map them directly to the correct structure. The syntax tree is also searched for questionable syntactic patterns to detect grammar errors that result in a syntactically valid sentence that is nonsensical in meaning.

For token-level errors, the checker reaches an error detection $F_{0.5}$ measure of 62.32, with typos reaching 92.66. For grammar errors, we currently reach an error detection $F_{0.5}$ measure of 24.64.

The spelling and grammar checker has been integrated into the editorial environment of an international CMS provider used by many Icelandic companies, including large media companies. Collaboration with a media company was used to carry out user tests and improve the user experience. The results show the checker to be a beneficial addition to the workflow.

To get the most usable and complete product for the largest user group by the end of the LTPI, the focus is on incorporating a neural language model, in particular more extensive coverage of grammar errors, error correction in general, and more detailed guidance tailored to different user groups.

3.5 Automatic Speech Recognition

The emphasis of the Automatic Speech Recognition (ASR) project within the LTPI has been on data collection, publication of quality ASR recipes and ultimately providing support for commercial applications depending on ASR. The following **data collections** have been ongoing during the project:

- **Read prompts** have been collected using the **Samrómur** (Mollberg et al., 2020) crowd-sourcing platform. The platform is derived from Mozilla’s Common Voice project⁷. The organization of the effort is based on the experience of a previous data collection efforts called **Málrómur** (Guðnason et al., 2012; Steingrímsson et al., 2017) and a platform called *Eyra* (Petursson et al., 2016). The web-based implementation of the platform has enabled easier organization of targeted collection

⁶<https://yfirlestur.is/>

⁷<https://commonvoice.mozilla.org/en>

efforts such as competitions aimed at children, teenagers and people who speak Icelandic as a second language. The platform uses a crowd-sourced verification system where users can vote for the correctness of read prompts. An automatic system based on forced-alignment scores (Guðnason et al., 2017) has been used to prioritize this effort. At the time of writing, 4,100 hours have been collected through this system in over 1,000,000 utterances. A part of the corpus has been published on OpenSLR (Mollberg et al., 2020) as well as on CLARIN-IS (Mollberg et al., 2021).

- **Broadcast news** corpus has been collected with the aid of The Icelandic National Broadcasting Service and CreditInfo's news watch service. The post-processing of this corpus has been extensive as the text needs to be aligned to the speech recordings to get a better time-resolution in the recorded segments. Force-alignment tools developed in-house (Guðnason et al., 2017) and the Montreal Forced Aligner (McAuliffe et al., 2017) have been used to create a database suitable for speech recognition training. At the time of writing, 487 hours have been collected through this system.
- **Question Answering** data set was collected using a specialized version of the **Samrómur** platform where the prompts were provided especially as questions. The prompts were obtained by pulling questions from the Icelandic Gigaword Corpus that fit certain criteria. In total, the data consists of 28 hours of recordings, of which 20 hours have been validated and published (Hedström et al., 2021).
- **Recorded lectures** were obtained from nine university lectures. Over 51 hours were transcribed by hand and published on CLARIN (Ragnarsson et al., 2022).
- **Speech dialogue** was obtained from conversations recorded on a specially created platform for two-ways conversations. On the platform a speaker is able to create a virtual chat room and share it with another speaker. The owner of the chat room can record the conversation and submit it. Submissions are then transcribed manually. In total, 21 hours have been collected and transcribed using this platform.
- **Other aligned recordings** have been collected and prepared for ASR before the start of the LTPI project with the collection of 542 hours of parliament speeches (Helgadóttir et al., 2017). Other publicly available sources are being explored. These include open court proceedings and rulings, recorded and transcribed municipality meetings and public parliament's committee meetings.

The utility of developing ASR recipes and applications alongside the data collection efforts is twofold. The obvious one is to create the technology that the data collections are intended to support. It therefore contributes directly to the main aims of the LTPI. The second utility is also very important which is to support and hone the data collection efforts with continuous feedback of quality and efficiency. The ASR recipe developers have had direct say in how the data is collected and curated; and they supported the post-processing of the data with forced time-alignment tools and automatic quality assessments. The **recipes and applications** developed during the project are:

- **ASR Kaldi Recipes** were developed for adult, adolescent, and children voices (Hernández Mena and Guðnason, 2022b; Hernández Mena and Guðnason, 2022a). The recipes were based on the read prompts from the Samrómur and Málrómur data collections (Helgadóttir et al., 2019; Nikulásdóttir et al., 2018b).
- **Web interface for ASR** was set up for both real-time speech streaming and off-line speech file uploads. The repository for setting this up was published with an open-source licence and the service runs online (Ragnarsson, 2021).
- **On-device ASR for smartphones** is in development for the Android operating system, using the Android Speech Recognition Service. The first version is scheduled for release by the end of this year.

- **ASR recipes** for voice control and question answering are being developed with a focus on a few specific tasks. This includes a specialized language model for questions.
- **Specialised ASR acoustic models** adapted to children, adolescents and people who speak of Icelandic as a second language.
- **Punctuation Prediction System** was developed for Icelandic using three different approaches: a BERT-based Transformer, a seq2seq Transformer and a bidirectional RNN.
- **Subword Unit Language Model** was developed for Icelandic which showed an improvement in word error rate.
- **Speaker Diarization Toolkit** for Icelandic using Kaldi (Fong and Guðnason, 2021).

All these **speech data collections** and **speech data recipes and tools** are being prepared for publication on CLARIN-IS.

3.6 Speech Synthesis

The focus of this core project has been on gathering sufficient resources and tools that are critical in developing a state-of-the-art text-to-speech system (TTS). The following resources are currently available on CLARIN-IS:

- **Talrómur** (Sigurgeirsson et al., 2021b) is a corpus containing 213 hours of speech recordings from eight different speakers. The corpus consists of four male voices and four female voices. The voices range in age, from 26 to 71 years old, and speaking style. In total, the corpus is made up of 122,417 single sentence utterances. The reading script was generated to maximize coverage of diphones in the Icelandic language and consists of sentences from multiple different sources (Sigurgeirsson et al., 2020; Sigurgeirsson et al., 2021a). The recordings were conducted in 2020 by Reykjavik University and RÚV, the Icelandic National Broadcasting Service, in a professional studio at the headquarters of the latter. To take the northern dialect in Icelandic into account, two of the voices were recruited from the north of Iceland and were recorded in a studio at the University of Akureyri. The audio is published in a single-channel 16-bit PCM wave file with a sample rate of 22050 Hz. Recordings were made using the recording platform LOBE (Sigurgeirsson et al., 2020) specifically designed for this purpose.
- **Talrómur 2** (Gunnarsson et al., 2021), is similar to Talrómur in many ways. It includes 80 hours of recordings from 40 different speakers with an even split of female and male voices. The voices were chosen to create four cohorts where each group consists of speakers with similar voice characteristics. Recordings were conducted in the same studio and with the same equipment as Talrómur. Both corpora share the same structure, format and audio specifications.
- Resources for TTS text pre-processing are i) a **text normalization corpus** (Sigurðardóttir, 2021) containing 140,000 sentences in their original form and automatically normalized for TTS (e.g. digits converted to their written-out forms and abbreviations expanded) and 40,000 manually normalized sentences, and ii) a **pronunciation dictionary** (Nikulásdóttir et al., 2018a; Nikulásdóttir et al., 2022) with around 65,000 manually verified entries, covering the four main pronunciation variants in Icelandic.
- Based on the text normalization corpus and the pronunciation dictionary, tools and models for TTS text pre-processing have been developed. For **automatic grapheme-to-phoneme conversion** (g2p) two approaches have been implemented: a rule-based module (Nikulásdóttir et al., 2020a), which is useful in lower resource settings, like on smartphones, and LSTM-based models. There is one model for each of the four pronunciation variants, and one model trained on Icelandic transcriptions of English words (Nikulásdóttir, 2020; Ármannsson, 2021). The **text normalization system** is based on regular expressions (Sigurðardóttir et al., 2021) and handles most cases of text normalization tasks

that can be expected in common news and sports texts. Modules for text cleaning, text normalization, adaptation of the spell and grammar checker, phrasing, and a complete g2p module including language detection, syllabification and stress labelling are already available on GitHub and will be published on CLARIN-IS in 2022. These modules build a complete TTS text-preprocessing pipeline, also to be published on CLARIN-IS.

- The evaluation platform **MOSI** (Jónsson et al., 2022) was created and is publicly available. MOSI supports multiple evaluation methods used for TTS systems, including MOS tests and A/B tests.

Training and development of TTS models using the Talrómur and Talrómur 2 corpora is underway. Tacotron2 and FastSpeech2 models have been trained on voices in Talrómur using the ESPNet toolkit, using phoneme inputs. Additionally, a parallel WaveGAN model has been trained on the entire Talrómur dataset. Evaluations are performed using MOSI. The models and evaluation results will be published by the end of the LTPI. The first models are already in use by a smartphone application developed within the LTPI, which will also be published on CLARIN-IS by the end of the programme.

4 Standards and Licencing

One of the core pillars of the LTPI is the publication of data and software under open licences. The guiding licences are CC BY 4.0⁸ for data and Apache 2.0⁹ for software. In exceptional cases, data have to be published with more restrictive licences, but all deliverables of the programme will be available for research and commercial use. An important part of ensuring open licensing is the crafting of agreements and consent statements for various data collection efforts.

All teams operate by common standards, defined in guidelines for data deliverables, on the one hand, and for software deliverables, on the other. Wherever possible, the guidelines adhere to international standards, e.g. regarding data format, metadata, or coding guidelines. Published data adhere to the FAIR standard¹⁰. Naming, versioning and keyword definitions are coordinated throughout the deliverables. Every software deliverable on CLARIN-IS has a link to the corresponding GitHub repository that is mostly hosted under the account of the developing partner. In general, the deliverables are separated modules, e.g. the tokenizer can be found as a stand-alone project. Other projects combine several modules, like the spell and grammar checker (see section 3.4) and text processing pipeline for TTS (see section 3.6).

Type of Repository	Number of Repositories
General text corpora, incl. test/dev	15
Specialized corpora	9
Parallel corpora	8
Lexical resources	12
NLP-tools	12
Machine translation	7
Spell and grammar checking	2
Speech corpora	6
Speech models and related modules	10
ALL REPOSITORIES	81

Table 2. CLARIN repositories from the LTPI. Status as of January 2022. Each project is only counted once but repositories have up to 3 previous versions, also available on CLARIN.

5 Usage Scenarios

The aim of the LTPI is that language resources and infrastructure software will be available for research and commercial use. The aimed-at users are LT specialists and general software developers that need to

⁸<http://creativecommons.org/licenses/by/4.0/>

⁹<https://www.apache.org/licenses/LICENSE-2.0>

¹⁰<https://www.go-fair.org/fair-principles/>

integrate LT in their products, as well as researchers from various fields.

There are numerous usage scenarios for the “buffet” of the LTPI deliverables. There are several levels of usage possibilities, reaching from low-level development using corpora and basic tools, to the usage of production-ready models or plugins/applications. For speech synthesis, for example, developers can use the speech corpora and necessary language-specific resources, like the pronunciation dictionary, to train and develop their own TTS models and voices. They can use the delivered TTS voices to integrate into their application, or they can use the web reader plugin directly to connect to their website.

As an example of products already using core resources, the full parser is the basis of the grammar checker within the LTPI. It is also used to parse questions and form answers for a voice assistant app, and is a module in an automatic term extraction software.

Table 2 shows the number of repositories on CLARIN-IS by January 2022 and how they can be divided into resource categories. We actively reach out to global players in LT to advertise the programme. In particular, we hope that carefully crafted, language-specific resources, like e.g. the TTS recordings and diverse gold and test corpora, will help lower the barrier for including Icelandic in existing global LT products.

It is also worth mentioning that the establishment of CLARIN-IS as the centre for Icelandic LT resources has led to Icelandic LT projects developed outside the programme to be published on CLARIN-IS as well. Thus, the foundation is laid for continuous delivery of LT resources to CLARIN-IS after the LTPI ends.

Acknowledgements

This project was funded by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almennarómur (<https://almannaromur.is/>), is funded by the Icelandic Ministry of Education, Science and Culture.

References

- Arnardóttir, Þ. and Ingason, A. K. 2020a. Icelandic Error Corpus Nonwords. CLARIN-IS, <http://hdl.handle.net/20.500.12537/63>.
- Arnardóttir, Þ. and Ingason, A. K. 2020b. nonwords. CLARIN-IS, <http://hdl.handle.net/20.500.12537/50>.
- Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., and Steingrímsson, S. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online), December. Association for Computational Linguistics.
- Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B., and Ingason, A. K. 2021. Creating an Error Corpus: Annotation and Applicability. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, pages 59–63, Virtual Edition.
- Ármannsson, B. 2021. Grapheme-to-Phoneme Transcription of English Words in Icelandic Text. Master’s thesis, Uppsala University.
- Barkarson, S. and Steingrímsson, S. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 140–145, Turku, Finland.
- Barkarson, S., Steingrímsson, S., Andrésdóttir, Þ. D., and Hafsteinsdóttir, H. 2020. IGC - Evaluation Set 20.09. CLARIN-IS, <http://hdl.handle.net/20.500.12537/51>.
- Barkarson, S., Andrésdóttir, Þ. D., Hafsteinsdóttir, H., Magnússon, Á. D., Rúnarsson, K., Steingrímsson, S., Jónsson, H. P., Loftsson, H., Sigurðsson, E. F., Rögnvaldsson, E., and Helgadóttir, S. 2021a. MIM-GOLD 21.05. CLARIN-IS, <http://hdl.handle.net/20.500.12537/113>.
- Barkarson, S., Steingrímsson, S., Ingimundarson, F. Á., Hafsteinsdóttir, H., and Magnússon, Á. D. 2021b. ParIce Dev/Test Sets 21.10. CLARIN-IS, <http://hdl.handle.net/20.500.12537/146>.

- Barkarson, S., Steingrímsson, S., and Hafsteinsdóttir, H. 2022. Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)* [to appear], Marseille, France.
- Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.
- Bjarnadóttir, K. and Ingimundarson, F. Á. 2021. DMII - Abbreviations 21.10. CLARIN-IS, <http://hdl.handle.net/20.500.12537/164>.
- Bjarnadóttir, K., Hlynsdóttir, K. I., and Steingrímsson, S. 2019a. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 146–154, Turku, Finland.
- Bjarnadóttir, K., Hlynsdóttir, K. I., Þórisson, S., Dagsson, T., and Steingrímsson, S. 2019b. DMII Core. CLARIN-IS, <http://hdl.handle.net/20.500.12537/12>.
- Bjarnadóttir, K. 2019. The Database of Modern Icelandic Inflection (DMII). CLARIN-IS, <http://hdl.handle.net/20.500.12537/5>.
- Bjarnadóttir, K. 2021. DIM Valency Structures 21.10. CLARIN-IS, <http://hdl.handle.net/20.500.12537/163>.
- Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.
- Daniélsson, H., Friðriksdóttir, S. R., and Steingrímsson, S. 2021. Icelandic Multi-SimLex (21.06). CLARIN-IS, <http://hdl.handle.net/20.500.12537/121>.
- Daniélsson, H., Jónsson, J. H., Árnason, Þ. A., Shaw, A., Sigurðsson, E. F., and Steingrímsson, S. 2021. The Icelandic Word Web: A Language Technology Focused Redesign of a Lexicosemantic Database. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 429–434, Reykjavík, Iceland.
- Daniélsson, H., Friðriksdóttir, S. R., Steingrímsson, S., and Sigurðsson, E. F. 2022. IceBATS: An Icelandic Adaptation of the Bigger Analogy Test Set. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)* [to appear], Marseille, France.
- Fong, J. Y. and Guðnason, J. 2021. RUV-DI Speaker Diarization (2021-10-14). CLARIN-IS, <http://hdl.handle.net/20.500.12537/157>.
- Friðriksdóttir, S. R., Daniélsson, H., and Steingrímsson, S. 2021. IceBATS - The Icelandic Bigger Analogy Test Set (21.06). CLARIN-IS, <http://hdl.handle.net/20.500.12537/120>.
- Friðriksdóttir, S. R. and Jasonarson, A. 2021. ALEXIA: Lexicon Acquisition Tool for Icelandic 3.0. CLARIN-IS, <http://hdl.handle.net/20.500.12537/123>.
- Friðriksdóttir, S. R., Jasonarson, A., Steingrímsson, S., and Sigurðsson, E. F. 2021. ALEXIA: A Lexicon Acquisition Tool. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, pages 64–67, Virtual Edition.
- Glišić, I. and Ingason, A. K. 2021. The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, pages 26–30, Virtual Edition.
- Guðjónsson, Á. A., Loftsson, H., and Daðason, J. F. 2021. Icelandic NER API – Ensemble Model. CLARIN-IS, <http://hdl.handle.net/20.500.12537/159>.
- Guðnason, J., Kjartansson, O., Jóhannsson, J., Carstensdóttir, E., Vilhjálmsón, H. H., Loftsson, H., Helgadóttir, S., Jóhannsdóttir, K. M., and Rögnvaldsson, E. 2012. Almannarómur: An Open Icelandic Speech Corpus. In *Spoken Language Technologies for Under-Resourced Languages*.
- Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. B. 2017. Building ASR Corpora Using Eyra. In *INTERSPEECH*, pages 2173–2177.
- Gunnarsson, Þ. D., Örnólfsson, G. T., Þórhallsdóttir, R., Sigurgeirsson, A. Þ., and Guðnason, J. 2021. Talrómur 2. CLARIN-IS, <http://hdl.handle.net/20.500.12537/165>.

- Hedström, S., Fong, J. Y., Þórhallsdóttir, R., Mollberg, D. E., Guðmundsson, S. F., Jónsson, Ó. H., Þorsteinsdóttir, S., Magnúsdóttir, E. H., and Guðnason, J. 2021. Samromur Queries 21.12. CLARIN-IS, <http://hdl.handle.net/20.500.12537/180>.
- Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. 2014. Correcting Errors in a New Gold Standard for Tagging Icelandic Text. In *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC)*, pages 2944–2948, Reykjavik, Iceland.
- Helgadóttir, I. R., Kjaran, R., Nikulásdóttir, A. B., and Guðnason, J. 2017. Building an ASR Corpus Using Althingi’s Parliamentary Speeches. In *INTERSPEECH*, pages 2163–2167.
- Helgadóttir, I. R., Nikulásdóttir, A. B., Borský, M., Fong, J. Y., Kjaran, R., and Guðnason, J. 2019. The Althingi ASR System. In *INTERSPEECH*, pages 3013–3017.
- Henrich, V., Reuter, T., and Loftsson, H. 2009. CombiTagger: A System for Developing Combined Taggers. In *Proceedings of the 22nd International FLAIRS Conference, Special Track: “Applied Natural Language Processing”*, Sanibel Island, Florida, USA.
- Hernández Mena, C. D. and Guðnason, J. 2022a. Icelandic Language Models with Pronunciations 22.01. CLARIN-IS, <http://hdl.handle.net/20.500.12537/172>.
- Hernández Mena, C. D. and Guðnason, J. 2022b. Samrómur-Children Demonstration Scripts 22.01. CLARIN-IS, <http://hdl.handle.net/20.500.12537/173>.
- Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. 2021a. The Icelandic Child Language Error Corpus (IceCLEC) version 1.1. CLARIN-IS, <http://hdl.handle.net/20.500.12537/133>.
- Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. 2021b. The Icelandic Dyslexia Error Corpus (IceDEC) version 1.1. CLARIN-IS, <http://hdl.handle.net/20.500.12537/132>.
- Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišić, I. 2021c. The Icelandic L2 Error Corpus (IceL2EC) version 1.2. CLARIN-IS, <http://hdl.handle.net/20.500.12537/131>.
- Ingólfssdóttir, S. L., Guðjónsson, Á. A., and Loftsson, H. 2020. Named Entity Recognition for Icelandic: Annotated Corpus and Models. In Espinosa-Anke, L., Martín-Vide, C., and Spasić, I., editors, *Statistical Language and Speech Processing*, pages 46–57, Cham. Springer International Publishing.
- Jónsson, H. P. and Loftsson, H. 2021. ABLTagger (Lemmatizer) – 3.1.0. CLARIN-IS, <http://hdl.handle.net/20.500.12537/134>.
- Jónsson, H. P., Loftsson, H., and Steingrímsson, S. 2020a. MT: Moses-SMT. CLARIN-IS, <http://hdl.handle.net/20.500.12537/46>.
- Jónsson, H. P., Símonarson, H. B., Snæbjarnarson, V., Steingrímsson, S., and Loftsson, H. 2020b. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In Sojka, P., Kopeček, I., Pala, K., and Horák, A., editors, *Text, Speech, and Dialogue*, pages 95–103, Cham. Springer International Publishing.
- Jónsson, J. H., Daníelsson, H., Árnason, Þ. A., and Shaw, A. 2020c. The Icelandic Wordweb 21.06. CLARIN-IS, <http://hdl.handle.net/20.500.12537/117>.
- Jónsson, H. P., Snæbjarnarson, V., Símonarson, H. B., and Þorsteinsson, V. 2021. En-Is Synthetic Parallel Named Entity Robustness Corpus. CLARIN-IS, <http://hdl.handle.net/20.500.12537/129>.
- Jónsson, S., Gunnarsson, Þ., Örnólfsson, G., and Sigurgeirsson, A. 2022. MOSI: TTS Evaluation Tool. CLARIN-IS, <http://hdl.handle.net/20.500.12537/186>.
- Jónsson, H. P., Loftsson, H., and Steingrímsson, S. 2021. ABLTagger 2.0. CLARIN-IS, <http://hdl.handle.net/20.500.12537/98>.
- Loftsson, H. and Östling, R. 2013. Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa)*, pages 105–119, Oslo, Norway.
- Loftsson, H. and Rögnvaldsson, E. 2007. IceParser: An Incremental Finite-State Parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NoDaLiDa)*, pages 128–135, Tartu, Estonia.
- Loftsson, H., Rögnvaldsson, E., and Pálsson, G. 2021. IceParser 1.5.0. CLARIN-IS, <http://hdl.handle.net/20.500.12537/122>.

- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *INTERSPEECH*, pages 498–502.
- Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Guðnason, J. 2020. Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 3463–3467, Marseille, France.
- Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Guðmundsdóttir, J., Steingrímsson, S., Magnúsdóttir, E. H., Fong, J., Borský, M., and Guðnason, J. 2021. Samromur 21.05. CLARIN-IS, <http://hdl.handle.net/20.500.12537/189>.
- Nikulásdóttir, A. B., Guðnason, J., and Rögnvaldsson, E. 2018a. An Icelandic Pronunciation Dictionary for TTS. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 339–345. IEEE.
- Nikulásdóttir, A. B., Helgadóttir, I. R., Pétursson, M., and Guðnason, J. 2018b. Open ASR for Icelandic: Resources and a baseline system. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Myasaki, Japan.
- Nikulásdóttir, A. B., Ármannsson, B., and Schnell, D. 2020a. Rule-based G2P for Icelandic. CLARIN-IS, <http://hdl.handle.net/20.500.12537/83>.
- Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. 2020b. Language Technology Programme for Icelandic 2019–2023. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France.
- Nikulásdóttir, A. B., Ármannsson, B., and Bryndís, B. 2022. Icelandic Pronunciation Dictionary for Language Technology 22.01. CLARIN-IS, <http://hdl.handle.net/20.500.12537/181>.
- Nikulásdóttir, A. B. 2020. Models for Automatic G2P for Icelandic. CLARIN-IS, <http://hdl.handle.net/20.500.12537/84>.
- Petursson, M., Klüpfel, S., and Guðnason, J. 2016. Eyra – Speech Data Acquisition System for Many Languages. *Procedia Computer Science*, 81:53–60.
- Ragnarsson, R. K., Mollberg, D. E., and Magnúsdóttir, E. H. 2022. Kennslurómur 22.01. CLARIN-IS, <http://hdl.handle.net/20.500.12537/171>.
- Ragnarsson, R. K. 2021. Tiro Web Interface for Speech Recognition 1.0. CLARIN-IS, <http://hdl.handle.net/20.500.12537/161>.
- Rúnarsson, K., Jónsson, B., and Gíslason, M. 2020. Icelandic Hyphenation Dictionary. CLARIN-IS, <http://hdl.handle.net/20.500.12537/86>.
- Rúnarsson, K. 2020. Skiptir. CLARIN-IS, <http://hdl.handle.net/20.500.12537/87>.
- Sigurðardóttir, H. S., Nikulásdóttir, A. B., and Guðnason, J. 2021. Creating Data in Icelandic for Text Normalization. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 404–412.
- Sigurðardóttir, H. S. 2021. Text Normalization Corpus 21.10. CLARIN-IS, <http://hdl.handle.net/20.500.12537/158>.
- Sigurgeirsson, A. Þ., Örnólfsson, G. T., and Guðnason, J. 2020. Manual Speech Synthesis Data Acquisition - From Script Design to Recording Speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 316–320, Marseille, France. European Language Resources association.
- Sigurgeirsson, A., Gunnarsson, Þ., Örnólfsson, G. T., Magnúsdóttir, E. H., Þórhallsdóttir, R. K., Jónsson, S., and Guðnason, J. 2021a. Talrómur: A Large Icelandic TTS Corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 440–444, Reykjavík, Iceland.
- Sigurgeirsson, A. Þ., Gunnarsson, Þ. D., Örnólfsson, G. T., Þórhallsdóttir, R., Magnúsdóttir, E. H., and Guðnason, J. 2021b. Talrómur. CLARIN-IS, <http://hdl.handle.net/20.500.12537/104>.
- Símonarson, H. B. and Snæbjarnarson, V. 2021. Icelandic Parallel Abstracts Corpus. *CoRR*, abs/2108.05289.
- Símonarson, H. B., Snæbjarnarson, V., and Þorsteinsson, V. 2020. En-Is Synthetic Parallel Corpus. CLARIN-IS, <http://hdl.handle.net/20.500.12537/70>.

- Símonarson, H. B., Snæbjarnarson, V., and Þorsteinsson, V. 2021a. En-Is Synthetic Parallel Corpus - 2021. CLARIN-IS, <http://hdl.handle.net/20.500.12537/127>.
- Símonarson, H. B., Snæbjarnarson, V., Ragnarson, P. O., Jónsson, H., and Þorsteinsson, V. 2021b. Miðeind's WMT 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.
- Snæbjarnarson, V., Símonarson, H. B., and Þorsteinsson, V. 2020. GreynirT2T Serving - En-Is NMT Inference and Pre-trained Models. CLARIN-IS, <http://hdl.handle.net/20.500.12537/72>.
- Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Jónsson, H. P., Ingólfssdóttir, S. L., and Þorsteinsson, V. 2021. GreynirTranslate - mBART25 NMT Models for Translations between Icelandic and English. CLARIN-IS, <http://hdl.handle.net/20.500.12537/125>.
- Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfssdóttir, S. L., Jónsson, H. P., Þorsteinsson, V., and Einarsson, H. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. *CoRR*, abs/2201.05601.
- Sólmundsdóttir, A., Stefánsdóttir, L. B., and Ingason, A. K. 2021. IceTaboo: a Database of Contextually Inappropriate Words for Icelandic. In Monachini, M. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference*, Virtual Edition.
- Steingrímsson, S. and Barkarson, S. 2021. ParIce 21.10. CLARIN-IS, <http://hdl.handle.net/20.500.12537/145>.
- Steingrímsson, S., Guðnason, J., Helgadóttir, S., and Rögnvaldsson, E. 2017. Málrómur: A Manually Verified Corpus of Recorded Icelandic Speech. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 237–240, Gothenburg, Sweden.
- Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., and Guðnason, J. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan.
- Steingrímsson, S., Kárason, Ö., and Loftsson, H. 2019. Augmenting a BiLSTM tagger with a Morphological Lexicon and a Lexical Category Identification Step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 1161–1168, Varna, Bulgaria.
- Steingrímsson, S., O'Brien, L. J., Ingimundarson, F. Á., Magnússon, Á. D., Andrésdóttir, Þ. D., and Eiríksdóttir, I. G. 2021. English-Icelandic/Icelandic-English Glossary 21.09. CLARIN-IS, <http://hdl.handle.net/20.500.12537/144>.
- Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.
- Þorsteinsson, V. and Óladóttir, H. 2020. Icegrams (2020-09-30). CLARIN-IS, <http://hdl.handle.net/20.500.12537/80>.
- Þorsteinsson, V., Óladóttir, H., and Loftsson, H. 2019. A Wide-Coverage Context-Free Grammar for Icelandic and an Accompanying Parsing System. In *Proceedings of the International Conference on Recent Advances in Natural Language Proceedings (RANLP)*, pages 1397–1404, Varna, Bulgaria.
- Þorsteinsson, V., Óladóttir, H., Arnardóttir, Þ., and Þórðarson, S. 2021a. GreynirCorrect. CLARIN-IS, <http://hdl.handle.net/20.500.12537/148>.
- Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. 2021b. BinPackage. CLARIN-IS, <http://hdl.handle.net/20.500.12537/137>.
- Þorsteinsson, V., Óladóttir, H., and Þórðarson, S. 2021c. GreynirPackage. CLARIN-IS, <http://hdl.handle.net/20.500.12537/147>.
- Þorsteinsson, V., Óladóttir, H., Þórðarson, S., and Ragnarsson, P. O. 2021d. Tokenizer. CLARIN-IS, <http://hdl.handle.net/20.500.12537/136>.
- Þorsteinsson, V., Óladóttir, H., Þórðarson, S., Símonarson, H. B., and Ásgeirsdóttir, K. 2021e. GreynirCorpus. CLARIN-IS, <http://hdl.handle.net/20.500.12537/119>.
- Þórðarson, S. 2020a. cities_is2en (2020-09-28). CLARIN-IS, <http://hdl.handle.net/20.500.12537/66>.

Þórðarson, S. 2020b. isprep4cc (2020-09-28). CLARIN-IS, <http://hdl.handle.net/20.500.12537/59>.

Þórðarson, S. 2020c. isprep4isloc (2020-09-28). CLARIN-IS, <http://hdl.handle.net/20.500.12537/58>.

Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction

Andrius Utka

Vytautas Magnus University
Kaunas, Lithuania
andrius.utka@vdu.lt

Sigita Rackevičienė

Mykolas Romeris University
Vilnius, Lithuania
sigita.rackeviciene@mruni.eu

Liudmila Mockienė

Mykolas Romeris University
Vilnius, Lithuania
liudmila@mruni.eu

Aivaras Rokas

Vytautas Magnus University
Kaunas, Lithuania
aivaras.rokas@vdu.lt

Marius Laurinaitis

Mykolas Romeris University
Vilnius, Lithuania
laurinaitis@mruni.eu

Agnė Bielinskienė

Vytautas Magnus University
Kaunas, Lithuania
agne.bielinskiene@vdu.lt

Abstract

The paper aims at presenting English-Lithuanian corpora for bilingual term extraction (BiTE) in the cybersecurity domain within the framework of the project DVITAS. It is argued that a system of parallel, comparable, and training corpora for BiTE is particularly useful for less-resourced languages, as it allows efficiently to combine strengths and avoid weaknesses of comparable and parallel resources. A special focus is given to the availability of sources in the cybersecurity domain and issues related to copyright-protected publications, as well as the data curation performed for building the corpora and depositing them to CLARIN-LT repository.

1 Introduction

The model of combining several types of corpora has been chosen for the bilingual terminology extraction project DVITAS.¹ The aim of the project is to develop a methodology for automatic extraction of English and Lithuanian terms of a specialised domain from parallel and comparable corpora, as well as to create a publicly available bilingual termbase. Cybersecurity (CS) terminology has been chosen as a specialised domain for the project because of its particular relevance in today's digitalised world in which cybersecurity awareness and cyber hygiene skills are indispensable for every Internet user. The compiled termbase is believed to be valuable both for specialists of the domain and the general public, as well as drafters of legal and administrative documents, and translators.

The project aims at employing current deep learning terminology extraction methods. In 2020, the project team (Rokas et al., 2020) completed a pilot study on semi-supervised automatic extraction of Lithuanian CS terms from a Lithuanian monolingual corpus. A small-scale manually annotated dataset (66,706 word corpus with 1,258 annotated cybersecurity terms) was used as a training data. The pilot study was performed in several stages: firstly, various baseline LSTM and GRU networks were tested using the Adam optimiser and FastText embeddings; secondly, each of the best baseline LSTM and GRU networks were tested with various optimisers; and finally, the best model was compared with a model that has been trained using multilingual BERT embeddings (Rokas et al., 2020). The latter approach proved to be the most efficient: Bidirectional Long Short-Term Memory model (Bi-LSTM) using multilingual Bidirectional Encoder Representations from Transformers (BERT) embeddings reached F1 score of 78.6%.

The methodology used in the pilot study will be modified and tested on different configurations of neural networks taking into account the methods applied in related research. In studies by other scholars,

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

¹<https://klc.vdu.lt/dvitas/en>

who applied neural networks for term extraction as sequence labelling task and used larger annotated datasets, higher F1 score was achieved: e.g., Kucza et al. used a dataset with 78,567 annotated terms and with Bi-LSTM reached F1 score of 86.73% (Kucza et al., 2018). Other studies on sequence labeling tasks with multilingual BERT embeddings show that reduction of the number of languages to three in BERT models may help to achieve higher results compared with the ones achieved with multilingual BERT (Ulčar and Robnik-Šikonja, 2020). As deep neural network models achieve higher and higher F1 scores, they reveal their performance and effectiveness in the tasks they were trained for and prove their spot as one of the peak state-of-the-art approaches to terminology extraction.

Thus, we believe that more comprehensive training and testing data obtained from larger bilingual corpora will allow to improve the preliminary results. Precisely, that is the goal of the present paper - to present the motivation behind the idea of creating such a resource, as well as to present solutions to encountered problems and challenges.

2 Related Research

Bilingual/multilingual term extraction, which is widely used for terminographic purposes, is performed by using two types of corpora - parallel and comparable. These two types of corpora are distinguished by the nature of texts that are used to build them. A parallel corpus (bilingual or multilingual) is the one "that contains source texts and their translations", whereas "a comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness", which means that it should include "the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period" (McEnery and Xiao, 2007).

Term extraction from parallel corpora has been already applied for several decades (Kupiec, 1993). It is considered to be relatively easy, at least from a technical point of view, as in a parallel corpus, which typically consists of aligned sentences, source and target terms appear in the same aligned pair of sentences. Parallel corpora are particularly useful for translation studies as their analysis provides insights into various equivalence issues. They are also extensively used to develop machine translation (MT) systems and computer-assisted translation (CAT) tools like translation memories (TM) (McEnery and Xiao, 2007). Moreover, "specialized parallel corpora can be especially useful in domain-specific translation research" (McEnery and Xiao, 2007).

Lately, the importance of comparable data have been increasing, as more and more papers have appeared on term extraction from comparable corpora (Vintar, 2010; Delpuch et al., 2012; Gornostay et al., 2012; Aker et al., 2013; Chu et al., 2016). Notably, since 2008 a number of valuable research papers on the usage of comparable corpora for term extraction have been published in Proceedings of the Workshop on Building and Using Comparable Corpora (BUCC)²

As the extraction of data from comparable corpora is not that straightforward and accurate as from parallel corpora, scholars have applied a variety of data extraction methods or combinations thereof. For instance, Steingrímsson et al. suggested combining three different approaches for effective bitext extraction from comparable corpora, namely combining crosslingual information extraction (CLIR), contextualised embeddings, and word alignments; this method is particularly useful for low-resourced scenarios (Steingrímsson et al., 2021). Sanjanasri et al. used Apache Spark framework for mining bilingual word pairs from a comparable corpus (Sanjanasri et al., 2021). Vintar et al. suggested applying intersections of word embeddings for mining semantic relations from comparable corpora (Vintar et al., 2020). Huidrom et al. proposed using the web as a source for building a comparable corpus for a less-resourced language pair using the heuristic approach based on sentence-length information and a bilingual dictionary when such is available (Huidrom et al., 2021). Terryn et al. presented a new approach to monolingual and multilingual term annotation and automatic term extraction based on the gold standard (Terryn et al., 2020).

Researchers indicate several important advantages of using comparable data. First, term extraction from comparable corpora provides valuable terminological data as these data reflect the usage of termi-

²Workshop on Building and Using Comparable Corpora (BUCC) - <https://aclanthology.org/venues/bucc/>.

nology in original languages which is much more natural than the usage of terminology in translations that are inevitably influenced by source languages. McEnery et al. also highlight that "specialised comparable corpora are particularly helpful for highly domain-specific translation tasks" (McEnery and Xiao, 2007). Another important advantage of using comparable data is the possibility to include data sources of a much larger variety, as comparable data is not limited to translated resources, which might be scarce or lack diversity, especially in cases when both the original language and the translation language are not English (Alonso et al., 2012; Delpech et al., 2012; Goeuriot et al., 2009; Morin and Prochasson, 2011; Morin et al., 2011; Rivera et al., 2013; Terryn et al., 2020). Thus, building and using comparable corpora for under-resourced languages next to parallel corpora could be very important for the analysis of such languages. And finally, comparable corpora are less expensive to build than parallel corpora as text alignment is not needed for their compilation.

Therefore, some scholars have introduced the idea of combining comparable and parallel corpora to benefit from the advantages provided by both (McEnery and Xiao, 2007; Bernardini, 2011; Morin and Prochasson, 2011; Biel, 2016; Giampieri, 2018), yet some researchers concentrate solely on comparable corpora (Steyaert and Rigouts Terryn, 2019; Vintar et al., 2020).

Thus, there is enough evidence to assume that for an efficient bilingual terminology extraction for English and Lithuanian languages, we need to build a resource consisting of parallel and comparable corpora.

3 Cybersecurity Domain and Availability of the Sources

The analysis of the cybersecurity sources revealed that this domain is highly heterogeneous and encompasses diverse types of information accumulated in various discourses. Ideally, the cybersecurity corpora should be representative of the whole cybersecurity domain and its constituent genres of texts produced in various discourses.

Wall in his study on cybercrime distinguishes four main discourses relevant to the CS domain: legislative/administrative discourse, academic discourse, expert discourse and popular, emotional or layperson's discourse (Wall, 2007). Similarly, for our corpora we distinguished legal, administrative-informative, academic, and media discourses. We ascribed expert texts written by cybersecurity practitioners to the administrative-informative discourse. The sources of these discourses were investigated and assessed for compilation of the corpora. Two most important criteria of source assessment were their suitability for compilation of a comparable corpus and a parallel corpus and their availability.

Most sources were suitable for compilation of the comparable corpus, which consists of the original texts in English and Lithuanian. Meanwhile, the sources suitable for the parallel corpus (English original texts and their translations into Lithuanian) were much more sparse. More detailed description of suitability of the sources for the parallel and comparable corpora is given in Subsection 4.1.1.

Though there were numerous sources suitable for corpora compilation, not all of them were freely available. Documents produced by national and international legislative and administrative bodies are commonly accessible without any restrictions. Meanwhile, the access to academic publications is often restricted. Most relevant academic sources are published by major publishing companies and protected by intellectual property rights. As we had to ensure proper usage of these texts, we examined the legal framework related to copyright protection and text and data mining (TDM) activities, as well as possibilities to acquire permissions to reuse relevant copyright-protected publications for corpora compilation, data extraction and storage in CLARIN-LT repository.

For a long time TDM activities have faced conservative intellectual property protection and strict restrictions (small-scale use, no possibility to develop derivative products, etc.) on the usage of legally protected sources. This situation has hampered big data projects which are necessary for development of various AI applications. Therefore, numerous studies have appeared discussing the situation and necessary changes in legal frameworks (Rosati, 2018; Sag, 2019; Flynn et al., 2020).

In the US, the notion of fair use of copyrighted works has been questioned and reinterpreted in law courts, which have ruled that copying of copyright-protected works for TDM research purposes satisfies fair use criteria and is not an infringement. The lawsuits concerned the Google Book Search Project (the

cases *Authors Guild v. GOOGLE*, 2015; *Authors Guild v. HaithiTrust*, 2014)³.

In the UK, a TDM exception was included in the statutory amendments to copyright law which came into effect in 2014. Since 2014 Copyrights, Designs and Patents Act (Article 29A) has allowed performing TDM activities provided that the TDM practitioner has lawful access to the resource and that TDM activities are performed for non-commercial purposes⁴.

Until recently, the EU did not have a uniform legislation regarding copyright protection related to TDM. Thus, TDM activities were subject to national copyright legislation and exceptions applied to copyright protection. E.g. in the Lithuanian copyright law (Law on Copyright and Related Rights of the Republic of Lithuania, last amended in 2015) there were no exceptions to copyright protection concerning TDM activities; Article 22 on reproduction of the copyright-protected work for purposes of teaching or scientific research did not respond to modern needs of TDM activities as it emphasized the purpose of illustration, short works, and short extracts⁵.

In 2019, The Directive on Copyright and Related Rights in the Digital Single Market (2019) was adopted with the aim to uniform and modernise the EU copyright protection, adapting it to the implementation of new technologies. Article 3 of the Directive "Text and data mining for the purposes of scientific research" states that "Member States shall provide for an exception to the rights provided for in <...> of this Directive for reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access." The Article also states that "Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results."⁶

The Directive had to be transposed to national legislations in two years; in case of delays, the Directive had to be applied directly. In December, 2021, the Lithuanian Law on Copyright and Related Rights amended according to the Directive's provisions was submitted to the legislature for adoption.

However, the EU Directive has already provoked criticism and calls for further development. The major problem remains sharing research datasets. The new Directive exempts researchers performing TDM activities from the obligation to obtain authorisation from rightholders of texts; however, "corpora compiled on the basis of such exceptions cannot be freely shared, which in a long run may have serious consequences for Open Science and the functioning of research infrastructures such as CLARIN ERIC" (Kamocki et al., 2019). Thus, even if research activities are freed from requirement to obtain permission from rightholders, "knowledge transfer, citizen science and user innovation may paradoxically become more difficult, as they require sharing of data between various groups of stakeholders" (Kamocki et al., 2019). In order to prevent this, "it is important to remember that even if certain research activities are exempted from the rules of copyright, proper licensing is still necessary to efficiently and widely share the fruits of researchers' work" (Kamocki et al., 2019).

In our work, the above-discussed legal problems became the reality. While working on collection of necessary sources, we selected 20 books on cybersecurity written by researchers and practitioners of the field and published by various publishing houses. As all books were copyright-protected, we contacted the publishing houses inside and outside the EU (8 in all) in order to request a permission to use the books for the compilation of our corpora and the storage in the CLARIN's repository.

We have found out that almost all publishing houses ask to fill permission request forms or use per-

³Prof. William T. Fisher III. *Authors Guild v. GOOGLE, Inc.* <https://opencasebook.org/casebooks/493-copyright/resources/9.2.5-authors-guild-v-google-inc/>; Michael Risch. *Authors Guild, Inc. v. HaithiTrust* <https://opencasebook.org/casebooks/409-an-open-internet-law-casebook/resources/6.3.1-authors-guild-inc-v-hathitrust/>

⁴Copyrights, Designs and Patents Act <https://www.legislation.gov.uk/ukpga/1988/48/contents>; CLARIN Legal Information Platform <https://www.clarin.eu/content/clic-text-and-data-mining-tdm-exceptions-uk-and-france>

⁵Law on Copyright and Related Rights of the Republic of Lithuania (English translation) <https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/5f13b560b2b511e59010bea026bdb259?jfwid=32wf6i76>

⁶Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC <http://data.europa.eu/eli/dir/2019/790/oj>

mission request systems. The request forms are meant mostly for commercial requests to republish the content owned by publishers in other publications. Some publishing houses have special request forms for educational purposes or for authors who want to reuse certain content (e.g. images, charts, tables) in their academic dissertations/theses. However, none of the forms were suitable for data mining and sharing; therefore, we had to include extensive comments explaining the specificity of our case.

As our application procedure was the same as the one of commercial requests, we had to indicate exact number of pages we want to reuse; the reuse of the whole book was not possible. We also had to indicate the title of the publication in which the content would be reused, the publisher, the format of publication, print run/number of expected users/download forecast, and other details relevant to print and online publications. We had to assure that we would be able to make the material secure, so that it would be password-protected against illegal copying/downloading/distribution. In most cases the permission to reuse the material could be obtained for one edition of the publication or a maximum period of one year.

In the comment slots (where they were available) we explained that the texts of the books would be reused as datasets for a research project on machine learning and terminology extraction, the processed texts would be stored in CLARIN-LT repository and the access to them will be restricted to academic users only via authentication service with university logins.

Despite our detailed explanations of our research aims in the forms, as well as correspondence with the publishers, none of the publishing houses granted us free reuse of the requested book extracts. The charges for an extract ranged from 200 to 5,000 Eur. The permission to use texts of one of the books was rejected because the rights were held by the author, not by the publishing house. As we did not have funds allocated for this in our budget, none of the copyrighted books were included in our corpora.

Our experience reveals that publishers do not have special permission request options for reuse of texts as datasets for TDM activities for scientific purposes. In addition, the publishers are not familiar with CLARIN infrastructure, its policy, aims and functions. Therefore, corpora stored in CLARIN repositories have to comply with the same requirements as commercial publications.

Thus, our corpora do not include copyright-protected books on cybersecurity which would be very important for our terminology extraction research. We had to rely on the inclusion of rather large bulk of publicly available media texts into the comparable corpus. In order to ensure its use for scientific purposes, we provided access to the corpus only to academic users of CLARIN-LT repository.

4 Corpora System for Bilingual Terminology Extraction

Five CS corpora have been compiled for this project: a parallel corpus of English texts and their Lithuanian translations (approx. 1.4 million words), a comparable corpus composed of two subcorpora: original English texts and original Lithuanian texts (approx. 4 million words), and three training (gold standard) corpora (approx. 0.1 million words each). The system of corpora and a flowchart of BiTE is presented in Figure 1.

Two of the corpora, namely *English-Lithuanian Parallel CS Corpus*⁷ and *English-Lithuanian Comparable CS Corpus*⁸, have been deposited to CLARIN-LT repository⁹. The parallel corpus is accessible under the CLARIN public licence (PUB), while the comparable corpus under the CLARIN academic licence (ACA).

The next subsections will present two important aspects of these two resources, namely data curation and composition.

4.1 Data Curation

Data curation is a very important and time-consuming activity, which ensures the quality and endurance of any dataset. The process of data curation typically involves the following steps: 1) discovering data sources; 2) acquiring textual data; 3) cleaning, deduplicating and transforming of extracted data, and 4)

⁷<https://clarin.vdu.lt/xmlui/handle/20.500.11821/46>

⁸<https://clarin.vdu.lt/xmlui/handle/20.500.11821/47>

⁹<https://clarin.vdu.lt/xmlui/>

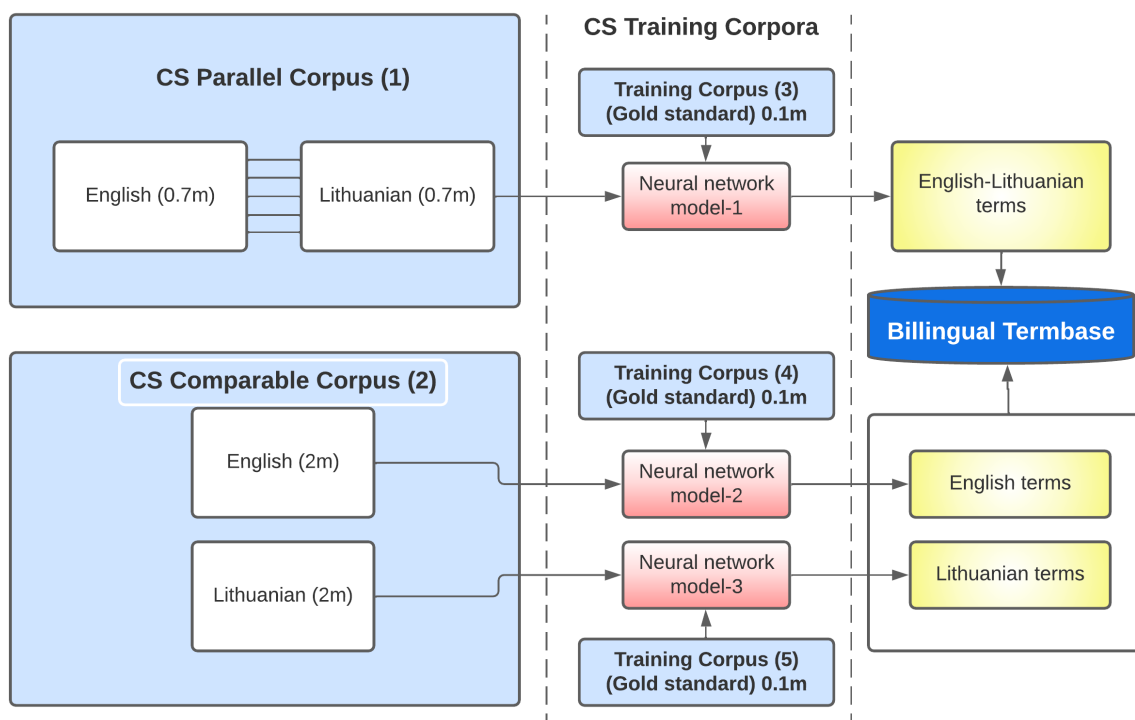


Figure 1. Corpora system for BiTE

integrating the data with other data sources. In the following subsections we will present the data curation steps, which are relevant for our project.

4.1.1 Discovering Data Sources

Presently, all the data for any corpora come from the web, however, texts from different discourses come in different quantities, formats and pose different challenges for a researcher.

First and most obvious source of textual information for our purposes is legal documents on cybersecurity, such as cybersecurity strategies, laws, government resolutions, minister orders, etc. Official national and EU legally binding and non-binding documents are commonly accessible without any restrictions. The documents of these categories can be acquired for both comparable and parallel corpora for both languages (see Table 4 and Table 3).

The second source of information is texts produced by CS experts (practitioners at national and international cybersecurity agencies and other institutions), containing reports, recommendations, information bulletins, guidelines, are also freely available; however, most of them are suitable only for the comparable corpus, as only a handful of them have been translated.

The third source of information is academic research publications on the cybersecurity topic. However, as we have shown in Section 3, access to academic publications is often restricted. Besides, research books and papers written in English and Lithuanian are seldom translated, thus, again acquired academic publications are only suitable for the comparable corpus.

The fourth source of information is media articles. This is by far the most voluminous source of information, as textual information on cybersecurity can be easily acquired by scraping various news portals ranging from general to specialised. Media articles can only be used for the comparable corpus, as only very rarely one can find genuine translations with alignable sentences.

As it could be expected, the volume of information on cybersecurity for English and Lithuanian lan-

guages differ across different sources. As mentioned previously, for the parallel corpus it is almost impossible to acquire original and translated academic texts and it is likewise difficult to find translated informative texts or media articles. Therefore the corpus relies mainly on EU English documents translated into Lithuanian.

The situation with the data for the comparable corpus is somewhat better: we could find data from all four discourses for both languages, but, clearly, Lithuanian sources cannot ensure the same diversity and be of a comparable size to global sources of the English language. Thus, the obvious solution was, firstly, to acquire as much as possible of Lithuanian data on cybersecurity, and then try to construct the similar structure for English.

4.1.2 Acquiring Textual Data

The data files containing relevant textual information have been acquired from the web by a variety of methods:

- using custom developed scrapers, benefiting from *Selenium WebDriver Beautiful Soup* modules for Python and targeting general portals (e.g. BBC for English and Delfi for Lithuanian), specific cybersecurity news portals (e.g. Bleeping Computer) and official EU portals (e.g. EUR-LEX); the method produces clean plain text files;
- manual downloading of PDF files, where possible (e.g. enisa¹⁰ portal);
- manual downloading of web-pages, where scraping was not practical;
- downloading of scientific works in PDF from our home universities' databases (e.g. master theses and doctoral dissertations);

Once the data files have been downloaded, the textual data need to be extracted from PDF, MS Word, or HTML files.

4.1.3 Cleaning, Deduplicating, and Transforming

The acquired data is not always intact:

- textual data extracted from PDF files often contain various problems with line breaking, extra spacing, extra tabs, processing pictures and tables, footnotes interfering with the main text, list of references, text in another language, etc.;
- scraped files are usually in a better shape, however, one can frequently download duplicates of the same text or extra information from a website; besides, dynamically loaded web pages may obscure the full data and introduce a plethora of web scraping issues.

The above mentioned problems are difficult to fix automatically, as cluttered files differ depending on a source. We had to use semi-automatic or even manual find-and-replace routines for cleaning the files. The deduplication process was alleviated by employing a custom fuzzy matching algorithm to the scraped data, which was then followed by a semi-manual checking.

After the files have been cleaned, they have to be transformed into the final form. For the parallel corpus it's semi-automatic alignment on the sentence level. We chose to use *LF Aligner*, which is well-suited for EU official documents (Varga et al., 2005). The resulting files are translation memory exchange (TMX) files.

In addition, English and Lithuanian texts of the comparable corpus have been morphologically annotated. For the English language we have used a large trained pipeline from the spaCy library¹¹. The resulting files are in a vertical tabulated format that marks "word", "lemma", "universal POS", and "fine grained POS"¹² (see Table 1). Due to it's minimal structural complexity, the chosen format is easily transformable into any other format.

¹⁰<https://www.enisa.europa.eu/>

¹¹<https://spacy.io/models>

¹²<https://github.com/explosion/spaCy>

A	a	DET	DT
wave	wave	NOUN	NN
of	of	ADP	IN
criticism	criticism	NOUN	NN
was	be	AUX	VBD
launched	launch	VERB	VRN
from	from	ADP	IN
the	the	DET	DT
privacy	privacy	NOUN	NN
supporters	supporter	NOUN	NNS
.	.	PUNCT	.
<s>	<s>		

Table 1. An example of an annotated sentence in the English CS comparable corpus

Likewise, the Lithuanian files have been morphologically annotated with a Lithuanian tagger from SEMANTIKA-2 project¹³, where the information of morphological analysis is presented as "word", "lemma", and "msd tag"¹⁴ (see Table 2).

Ekspertai	ekspertas	Ncmpnn-
vieningai	vieningai	Rgp
teigė	teigti	Vgma3—n-ni-
,	,	Tc
kad	kad	Cg
reiškinys	reiškinys	Ncmsnn-
pavojingas	pavojingas	Agpmsnn
.	.	Tp
<s>	<s>	Xh

Table 2. An example of an annotated sentence in the Lithuanian CS comparable corpus

4.1.4 Integrating with Other Data Sources

In our case the integration with other data sources has involved the storage and sharing of the compiled corpora on the CLARIN-LT repository. CLARIN's DSpace-based depositing service (Mišutka et al., 2015) is conveniently built and as a depositor you only will be required to consider the following steps:

- description of the data resource;
- supplying the resource with relevant metadata;
- choosing appropriate format acknowledged by the research community;
- choosing of an appropriate licence;
- archiving of the data.

4.2 Composition of English-Lithuanian CS Corpora

4.2.1 English-Lithuanian Parallel CS Corpus

The parallel corpus includes the EU legal acts and other documents from the time period of 2010-2020. There are 80 files in English and Lithuanian aligned on the sentence level in the corpus. The total size is 1.4m words (EN - 773,373; LT - 633,942). The number of unique words (*types*) is 12,171 for English,

¹³<https://semantika.lt/>

¹⁴<https://github.com/Semantika2/Morfologiniu-zymeliu-standartas>

and 31,558 for Lithuanian. The corpus contains 35,415 aligned segments. The documents are extracted from the EUR-LEX database and other EU institutional repositories (see Table 3).

Document categories	Subcategories	Proportion
Legally binding (secondary legislation)	Regulations of the European Parliament and of the Council; Directives of the European Parliament and of the Council; Decisions of the European Parliament and of the Council	60%
Official non-binding	Communications of the European Commission; Reports of the European Commission; Recommendations of the European Commission; Opinions of the Committees of the EU; Briefing papers of the Court of Auditors	40%

Table 3. Structure of the parallel corpus (2010-2020)

4.2.2 English-Lithuanian Comparable CS Corpus

The CS comparable corpus compiled for the project includes texts from the time period of 2010-2021 (except for a few important documents from an earlier period). There are 1,708 files in English and 2,567 in Lithuanian. The total size of the corpus is 4m words (EN - 2,000,586; LT - 2,000,343). The number of unique words (*types*) is 37,565 for English, and 101,076 for Lithuanian. Text categories, subcategories and their proportions within the corpus are presented in Table 4.

Text categories	Subcategories	EN	LT
Academic	Scientific articles, monographs, MA and PhD theses, textbooks	19%	30%
Administrative-informative	Reports and recommendations of Cybersecurity Centres; booklets and posters	8%	11%
Legal	CS strategies, laws, government resolutions, ministry orders	18%	4%
Media	Mass media articles, specialised media articles	55%	55%

Table 4. Structure of the comparable corpus (2010-2021)

When compiling a comparable corpus it is very important to ensure similar sampling procedures for compared languages, as the goal is to compare how a particular domain is reflected in two distinct languages. As mentioned earlier, in the case of English-Lithuanian CS comparable corpus, it was difficult to attain the ideal balance of text categories (see Figure 2). Nevertheless, we have achieved, that the media part and other parts (legal, administrative-informative and academic) if taken together, would be equal for both languages (55% and 45%). Thus, we have attained the balance between two parts of the corpus, one of which is more popularised, while the other is more specialised.

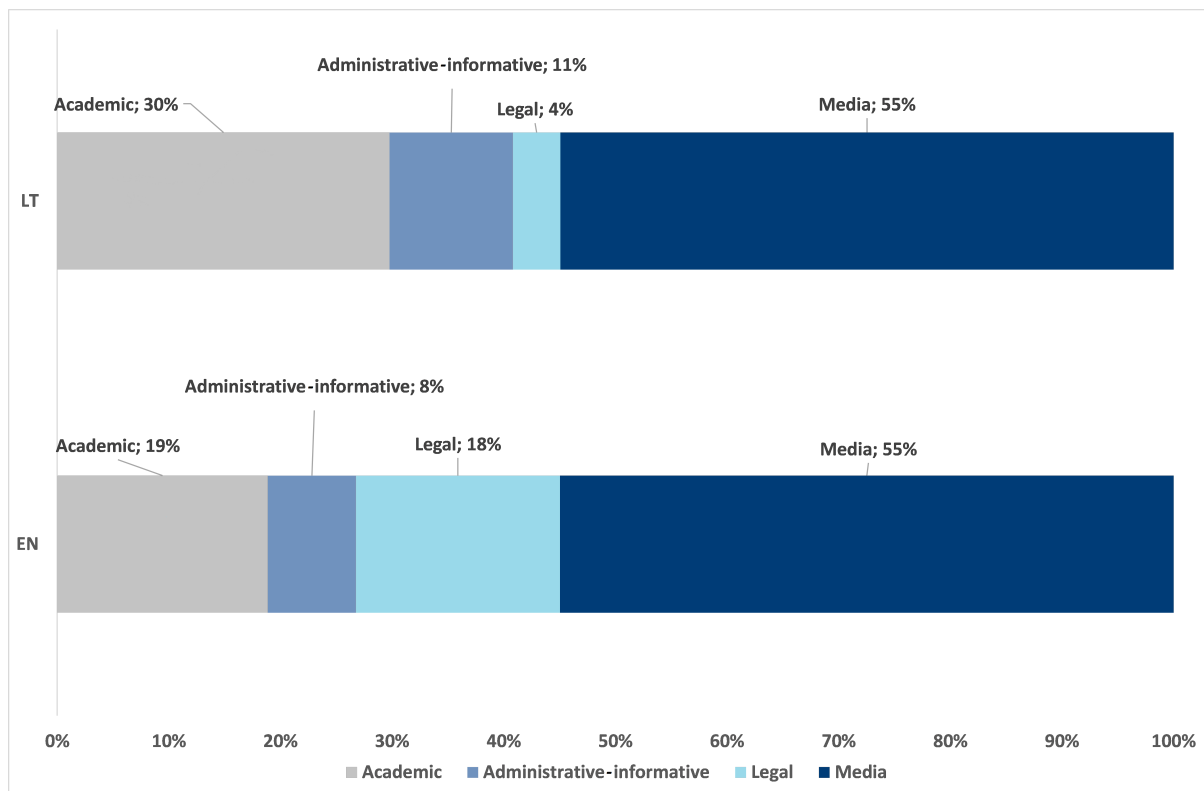


Figure 2. Subcorpora proportions in the English-Lithuanian CS comparable corpus

Ten most frequent lemmas of English and Lithuanian common nouns suggest that the two subcorpora in the comparable corpus are comparable, as 7 out of 10 nouns are at the top 10 in the both corpora:

1. security	14,825	saugumas	'security'	23,110
2. information	7,724	duomuo	'datum'	14,357
3. attack	7,338	sistema	'system'	11,910
4. system	6,384	informacija	'information'	10,815
5. datum	6,175	tinklas	'network'	7,817
6. cybersecurity	6,073	internetas	'internet'	7,061
7. threat	4,613	virtotojas	'user'	6,760
8. network	4,536	valstybė	'state'	6,479
9. service	4,129	programa	'program'	6,169
10. user	3,855	ataka	'attack'	6,121

4.2.3 Training corpora

In order to train neural networks to perform BiTE, training corpora (gold standard) have been compiled. They have been composed of the same text categories as the main corpora. The comparable training corpora contain legal texts (legislative acts and government resolutions), administrative-informative texts (reports and recommendations by CS experts), academic publications (theses and textbooks), and media articles. Parallel training corpus is composed of the most important EU legal acts and other documents on cybersecurity issues.

The corpora are being manually annotated by tagging three categories of terminological data: terms of the CS domain, terms related to the CS domain, as well as proper names relevant to the CS domain. Four terminology researchers are working on annotation of the training corpora in constant cooperation with a cybersecurity expert who consults and validates the annotation results (see more in (Rackevičienė

et al., 2021)).

5 Concluding Remarks

The analysis of cybersecurity sources revealed that this domain is highly heterogeneous and encompasses diverse types of information accumulated in various discourses. However, availability of some sources is limited. The limitations mainly concern the scientific publications, most of which are copyright-protected. Their reuse for TDM activities and storage in research data repositories involves tackling complex (and not adapted to these aims) permission request procedures and often are charged by rightholders. Though CLARIN constantly raises legal issues related to storing and sharing language resources, further steps are evidently needed by the research community to foster the development of Open Science.

Acquisition of quality data and its curation proved to be a challenging task due to its dynamic nature, necessitating manual reviewing and removing clutter or fixing incomplete elements of the data. The data curation is time-consuming and never ending process, which, nevertheless, needs to be continued in order to ensure quality and longevity of a resource.

Despite the fact that we could not include all planned sources to our corpora, the compiled parallel and comparable corpora contain reasonable variation, as they represent the cybersecurity domain in four different discourses in international and national settings. Thus, we believe that the corpora will provide sufficient data for the future research: deep learning-based terminology extraction, terminology analyses, as well as compilation of a bilingual cybersecurity termbase.

Acknowledgements

The research is carried out under the project “Bilingual Automatic Terminology Extraction” funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282). The project is also included as a use case in COST action “European Network for Web-Centred Linguistic Data Science” (CA18209).

References

- Aker, A., Paramita, M. L., and Gaizauskas, R. J. 2013. Extracting bilingual terminologies from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 402–411. The Association for Computer Linguistics.
- Alonso, A., Blancafort, H., de Groc, C., Million, C., and Williams, G. 2012. Metricc: Harnessing comparable corpora for multilingual lexicon development. In *15th EURALEX International Congress*, pages 389–403.
- Bernardini, S. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS - A Journal of Professional Communication*, 26.
- Biel, Ł. 2016. Mixed corpus design for researching the eurolect: A genre-based comparable-parallel corpus in the pl eurolect project. *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*.
- Chu, C., Dabre, R., and Kurohashi, S. 2016. Parallel sentence extraction from comparable corpora with neural network features. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).
- Delpech, E., Daille, B., Morin, E., and Lemaire, C. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In Kay, M. and Boitet, C., editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 745–762. Indian Institute of Technology Bombay.
- Flynn, S., Geiger, C., Quintais, P. J., Margoni, T., Sag, M., Guibault, L., and Carroll, M. W. 2020. Implementing user rights for research in the field of artificial intelligence: A call for international action. *European Intellectual Property Review*, 42(7):393–398.
- Giampieri, P. 2018. Online parallel and comparable corpora for legal translations. *Altre Modernità*, 20:237–252.

- Goeuriot, L., Morin, E., and Daille, B. 2009. Compilation of specialized comparable corpora in french and japanese. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC 2020)*, pages 55–63.
- Gornostay, T., Ramm, A., Heid, U., Morin, E., Harastani, R., and Planas, E. 2012. Terminology extraction from comparable corpora for latvian. In Tavast, A., Muischnek, K., and Koit, M., editors, *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012, Tartu, Estonia, 4-5 October 2012*, volume 247 of *Frontiers in Artificial Intelligence and Applications*, pages 66–73. IOS Press.
- Huidrom, R., Lepage, Y., and Khomdram, K. 2021. Em corpus: a comparable corpus for a less-resourced language pair manipuri-english. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67.
- Kamocki, P., Ketzan, E., Wildgans, J., and Witt, A. 2019. New exceptions for text and data mining and their possible impact on the clarin infrastructure. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, Linköping Electronic Conference Proceedings, pages 66–71. Linköping University Electronic Press, Linköpings universitet.
- Kucza, M., Niehues, J., Zenkel, T., Waibel, A., and Stüker, S. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In Yegnanarayana, B., editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 2072–2076. ISCA.
- Kupiec, J. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 17–22.
- McEnery, A. and Xiao, Z. 2007. Parallel and comparable corpora: What is happening? In Anderman, G. and Rogers, M., editors, *Incorporating Corpora: The Linguist and the Translator*, pages 18–31. Multilingual Matters.
- Mišutka, J., Kamran, A., Košarko, O., Josifko, M., Ramasamy, L., Straňák, P., and Hajič, J. 2015. Linguistic digital repository based on DSpace 5.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Morin, E. and Prochasson, E. 2011. Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (BUCC 2011)*, pages 27–34.
- Morin, E., Hazem, A., and Saldarriaga, S. P. 2011. Bilingual lexicon extraction from comparable corpora as metasearch. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web (BUCC 2011)*, pages 35–43.
- Rackevičienė, S., Utkā, A., Mockienė, L., and Rokas, A. 2021. Methodological framework for the development of an english-lithuanian cybersecurity termbase. *Studies about Languages/Kalbų studijos*, 39:85–92.
- Rivera, O. M., Mítkov, R., and Pastor, G. C. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. *Multiword Units in Machine Translation and Translation Technology*.
- Rokas, A., Rackevičienė, S., and Utkā, A. 2020. Automatic extraction of lithuanian cybersecurity terms using deep learning approaches. In Utkā, A., Vaičėnėnienė, J., Kovalevskaitė, J., and Kalinauskaitė, D., editors, *Human language technologies - the Baltic perspective: proceedings of the 9th international conference, Baltic HLT, Kaunas, Vytautas Magnus University, Lithuania, 22-23 September 2020*, pages 39–46. IOS Press.
- Rosati, E. 2018. *The exception for text and data mining (TDM) in the proposed Directive on Copyright in the Digital Single Market : technical aspects*. European Parliament.
- Sag, M. 2019. The new legal landscape for text mining and machine learning. *Journal of the Copyright Society of the USA*, 66.
- Sanjanasri, Menon, V. K., Kp, S., and Wolk, K. 2021. Mining bilingual word pairs from comparable corpus using apache spark framework. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 2–7.
- Steingrímsson, S., Lohar, P., Loftsson, H., and Way, A. 2021. Effective bitext extraction from comparable corpora using a combination of three different approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17.

- Steyaert, K. and Rigouts Terryn, A. 2019. Multilingual term extraction from comparable corpora: Informativeness of monolingual term extraction features. In Sharoff, S., Zweigenbaum, P., and Rapp, R., editors, *Proceedings of the 12th Workshop on Building and Using Comparable Corpora (BUCC 2019)*, pages 16–25, Varna, Bulgaria, September.
- Terryn, A. R., Hoste, V., and Lefever, E. 2020. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Language Resources and Evaluation*, 54(2):385–418.
- Ulčar, M. and Robnik-Šikonja, M. 2020. Finest bert and crosloengual bert: less is more in multilingual models. *arXiv e-prints*, June.
- Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596.
- Vintar, Š., Grčić Simeunović, L., Martinc, M., Pollak, S., and Stepišnik, U. 2020. Mining semantic relations from comparable corpora through intersections of word embeddings. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora (BUCC 2020)*, pages 29–34, Marseille, France, May. European Language Resources Association.
- Vintar, Š. 2010. Bilingual term recognition revisited. *Terminology*, 16:141–158.
- Wall, D. S. 2007. *Cybercrime: The Transformation of Crime in the Information Age*. Polity, Cambridge, UK.

‘Cretan Institutional Inscriptions’ Meets CLARIN-IT

Irene Vagionakis

University of Bologna, Italy
irene.vagionakis2@unibo.it

Paola Baroni

CNR-ILC – Pisa, Italy
paola.baroni@ilc.cnr.it

Riccardo Del Gratta

CNR-ILC – Pisa, Italy
delgratta@ilc.cnr.it

Angelo Mario Del Grosso

CNR-ILC – Pisa, Italy
angelo.delgrosso@ilc.cnr.it

Federico Boschetti

CNR-ILC & VeDPH – Pisa, Italy
federico.boschetti@ilc.cnr.it

Tiziana Mancinelli

VeDPH – Venezia, Italy
tiziana.mancinelli@unive.it

Monica Monachini

CNR-ILC – Pisa, Italy
monica.monachini@ilc.cnr.it

Abstract

This paper presents *Cretan Institutional Inscriptions*, a resource in the domain of Digital Epigraphy developed at the Ca’ Foscari University of Venice and supported by CLARIN-IT as part of its actions addressed to initiatives, projects and events in the field of Social Sciences and Humanities. The paper begins with a brief outline of the project within which the resource was created and then goes into a more in-depth description of the main methodologies used to develop the resource (EpiDoc and EFES) and of their benefits. The paper then focuses on the cooperation of the project with the Venice Centre of Digital and Public Humanities and the Italian node of CLARIN, also illustrating the dockerization process applied to the resource hosted on the CLARIN-IT servers. Some desiderata for future developments are outlined as well. The paper ends with some remarks about the widening of CLARIN horizons towards Digital Epigraphy and on the role of its K-Centres in this respect.

1 Project Description

The EpiDoc collection named *Cretan Institutional Inscriptions*¹ was created as a part of the Ph.D. research project in Ancient Heritage Studies *Kretikai Politeiai: Cretan Institutions from VII to I century BC*, carried out by Irene Vagionakis at the Ca’ Foscari University of Venice (UNIVE) from 2016 to 2019 under the supervision of Claudia Antonetti and Gabriel Bodard. The database, built by using the EpiDoc Front-End Services (EFES), collects the EpiDoc editions of 600 inscriptions shedding light on the institutions of the political entities of Crete from the VII to the I century BC.

The project, which contributes to the landscape of Digital Humanities – in particular to that of Digital Epigraphy, through the creation of a new open access online epigraphic resource – and could hopefully be a forerunner for the inclusion of other digital epigraphy projects in the Common Language Resources and Technology Infrastructure (CLARIN), has been a valuable opportunity for collaboration with the Venice Centre for Digital and Public Humanities (VeDPH) and the Italian node of CLARIN (CLARIN-IT) during its final testing and publication stages.

1.1 Aim of the Research

The Archaic, Classical and Hellenistic history of Crete is a history characterised by a very high level of fragmentation. The numerous silences of the literary sources and the gaps in the epigraphic records have resulted in wide sectors of the island history still being overshadowed and in a similar fate befalling

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

¹For a detailed description of the resource, see Vagionakis, 2021.

many of its political entities. In particular, the institutional history of Crete was greatly affected by such fragmentation and also by the bulky presence of the albeit scarce literary sources related to it. In fact, the alleged greater authoritativeness of authors such as Plato, Aristotle or Ephorus often prompted to force what was witnessed by the uneven epigraphic records to relate the contradictory information coming from different areas to a single model, thus flattening the variegations of the multiform landscape of «one hundred-cities Crete» (Hom. *Il.* II 649) in the name of the existence of a single unitary *Kretike politeia*. Within that framework, the Ph.D. research project set itself the goal of collecting systematically the records pertaining to Cretan institutions in order to propose an up-to-date reconstruction of the administrative framework of the island political entities, highlighting the specificity of each context, from the rise of the *poleis* and their first epigraphic records in the Greek alphabet(s) to the Roman conquest of Crete (VII-I century BC). By bringing together these so far scattered records in a searchable digital collection, the project also aimed at facilitating their finding, consultation and reuse.

2 The EpiDoc Collection and the TEI Catalogues

The core of the documentary basis of the research consisted of 600 Greek inscriptions, either directly mentioning institutional elements (as the decree from Knossos *I.Cret.* I 8 12 of the late II cent. BC: I. 1, ἔδοξε Κνωσίων τοῖς κόσμοις καὶ τῆι πόλι, ‘the *kosmoi* and the *polis* of the Knossians decreed’) or hinting at them through a revealing terminology (as the treaty between Hierapytna and an unknown *polis* *I.Cret.* III 3 6, of the late III or early II cent. BC: ll. 1–2, μηνὸς [- - -] τάδε ἔδ[οξ]εν τ[- - -], ‘in the month of [- - -], [- - -] decreed these things’).

For each inscription, an XML edition compliant with the TEI EpiDoc epigraphic substandard (Elliott et al., 2020) was created, including a descriptive and a bibliographic lemma, the text of the inscription, a selective apparatus criticus and a commentary focused on the institutional data offered by the document, plus links to other related online resources.

The EpiDoc markup was especially functional to the research questions in the encoding of the Greek texts inside `<div type="edition">`, where its semantic nature proved to be very helpful for the extraction of the institutional elements and for the analysis of variations in their type and function or sphere of competence, avoiding preconceived generalizations and valuing the specificity of each occurrence. In fact, the markup of the institutional elements was based on the use of `<rs type="institution">` along with a customized combination of focused attributes: `@subtype` for specifying their typology (such as assembly or board), `@role` for specifying their function or field of action (such as voter or dedicant), `@ref` for specifying their political entity (usually their *polis*), `@key` for facilitating their indexing through normalized forms of their names. A complete example of the markup of an institutional element can be found in Figure 1 below.

```
<rs type="institution" subtype="official" role="eponym" key="damiorgos" ref="#olous">ἐπὶ
<w lemma="δαμιοργός">δαμιοργοῦ</w> <persName type="attested"
key="Leukos"><name nymRef="Λεῦκος">Λεύκου</name></persName></rs>
```

Figure 1. An excerpt from the XML EpiDoc text of inscription *I.Cret.* I 22 4 A (l. 31)

In addition to `<rs>`, some other core TEI EpiDoc elements were used: `<w lemma="">` for the lemmatization of institutional and other relevant terms, `<placeName type="" ref="">` for toponyms and ethnic adjectives, `<persName type="" key="" ref="">` and `<name nymRef="" type="">` for prosopographic and onomastic elements (for officials, honoured individuals, foreign rulers and theonyms).

Overall, the semantic markup involved 8,162 lemmata (`<w>`), 4,353 institutional elements (`<rs>`), 2,633 toponyms or ethnic adjectives (`<placeName>`), 1,694 anthroponyms (`<name>`) and 1,651 prosopographical elements (`<persName>`).

In addition to the epigraphic collection, the research outputs also include the creation of two TEI catalogues: one relating to the political entities of Crete (*poleis, koina*, dependent communities, extra-urban sanctuaries and hegemonic alliances); another one relating to the attested Cretan institutions (comprising assemblies, boards, officials, associations, civic subdivisions, social statuses, age classes, months, festivities and other celebrations, institutional practices, institutional instruments and public spaces).

3 Benefits of Using EFES and EpiDoc

The EpiDoc Front-End Services (EFES) are an open source customisable tool for the online publication of ancient documents in EpiDoc XML, inscriptions in *primis* (Bodard and Yordanova, 2020)². It is the EpiDoc specialisation of Kiln, an analogous framework for publishing collections of TEI XML documents, from which it was forked in 2017³. The main strengths of EFES, as well as of its ancestor Kiln, are its comprehensiveness, ease of use and high customizability, which make it possible to quickly create a Web site provided with indices, textual search and browse facilities even from persons without advanced IT skills. In fact, the aim of EFES is to allow the production of such outputs also from small projects whose teams neither include IT experts devoted to the development of Web sites nor have funds to be dedicated to that purpose.

The specific case of *Cretan Institutional Inscriptions* is particularly emblematic of the benefits deriving from the use of EFES, being it the first project carried out by a single person to have used it. Despite the awareness of the importance and usefulness of a collaborative approach to research, the case was that of an individual doctoral project to be completed in three years with no external support, thus perfectly matching the target of users expected by EFES. The timing of the first release of EFES in September 2017, at the end of the first year of the Ph.D. research, was providential and allowed the creation of the Web site in the remaining two years.

In particular, with some customization of the provided XSLT stylesheets, from the EpiDoc markup of the inscriptions it was possible to generate several custom thematic indexes, recording the occurrences of institutional elements, relevant lemmas, prosopography, onomastics, toponyms, ethnics and theonyms. These indexes, especially the one relating to the institutions, are displayed in a tabular format, where all the pieces of information included in the markup are collected in separate columns and can be easily combined and compared with each other.

Besides the indexes, the EpiDoc encoding allowed the creation of very specific search filters, thanks to which the inscriptions can be browsed not only according to their traditional metadata (type of document, type of support, date, provenance, current location, bibliographic reference), but also on the basis of the name (e.g. *agela*, «herd»), type (e.g. tribe) and role (e.g. decreer) of the institutional elements and of the name of the places and divinities mentioned.

Another benefit deriving from the EpiDoc encoding is the high level of accuracy of the textual searches performed on the collection, which ignore all the extremely frequent diacritics due to the epigraphic editorial conventions and can be further refined by including the lemmatized base forms of the terms.

From a linguistic point of view, lemmatization proved to be a rather significant component of the work, considering the nature of the encoded inscriptions. In fact, the texts present themselves as particularly remarkable from a linguistic and dialectal perspective. The inscriptions, in the Cretan Doric dialect, contain a large number of lemmas – mostly terms attested only in epigraphic sources – that are either difficult to be found or completely absent in the main lexica of ancient Greek (such as the renowned Liddell-Scott-Jones Greek-English Lexicon; they are often mentioned within the corresponding Attic lemmas only)⁴, including some *hapax legomena* (such as *δωροτελέω*, *συνβολήτρα*) that are unique from

²EFES: code <https://github.com/EpiDoc/EFES>, documentation <https://github.com/EpiDoc/EFES/wiki>.

³Kiln: code <https://github.com/kcl-ddh/kiln>, documentation <https://kiln.readthedocs.io/en/latest>.

⁴Some examples among the many possible are *ἀγέλαος*, *ἀγρήιον*, *βοανθέω*, *γυνά*, *δαμοργός*, *δαρχνά*, *ἐσζικαιωτήρ*, *ἐσπράττω*, *κσενοδόγος*, *μνάμων*, *ματρῶια*, *νενομήια*, *οικετήια*, *πράδδα*, *σαλπίνδα*, *τριφετηρία*, *τριόδελον*, *ψαφίδδα*, *ώνά*. Curiously, the printed version of the Liddell-Scott-Jones as well as its online version plus other online lexica that can be accessed through the shared Logeion interface (<https://logeion.uchicago.edu>), do not even include *ποινικαστάς*, an element so emblematic of Archaic Cretan Doric dialect and culture that it was chosen as the name of the online resource

a lexical or morphological point of view.

Therefore, the lemmatization carried out was aimed at preserving the dialectal particularities of the Dorian Cretan language, including the presence of letters such as digamma (Ϝ) and qoppa (Ϙ), attested the latter also in alternation with kappa in cases such as ῥόσμος/κόσμος, ῥοσμῆω/κοσμῆω, ὄρῥος/ὄρκος, πρόῥοος/πρόκοος.

In addition to the linguistic peculiarities, the epigraphic nature of the database is also an aspect that made lemmatization particularly useful. In fact, the high fragmentary nature of the texts led to the presence of very frequent diacritical marks, particularly in relation to the indication of lacunae, additions and uncertain readings (square brackets, dashes, dots, underdots, question marks; see e.g. the text in Figure 2), which severely limits the search possibilities using the simple text-by-string search method. The combined action of EpiDoc markup and lemmatization made it possible to ignore diacritical marks in text searches as well as to perform text searches based on the lemmatized text.

35. Iscrizione edificatoria degli *eunomiotai* di Aptaera

Tipologia documentaria: iscrizione edificatoria

Supporto: sconosciuto

Datazione: II secolo a.C.

Provenienza: Aptaera

Collocazione attuale: iscrizione probabilmente perduta

Edd. Haussoullier 1879, p. 436, n. 10; *SGDI* 4949; Guarducci 1933, n. 5; *IG II 3 21* ✓**PHI**.

[-----]
[-- -]ν Εὐρυμῆδης Ἀνδι[-- -],
[-- -]χος Ἀρχέτω, Ὀρσικλῆ[ς -- -],
[-- -]σκος Ὀξινμ[άχ(?)ω],
[-- -] Ἀ]λκιμένη ἐπεμελήθη[ν -- -]
5 [-- -]ρσιος καὶ τᾶν λοιπᾶν πα[σᾶν -- -]
[-- -] μέστᾳ ἐπὶ τ[.] εὐνομιῶτ[αν -- -]
[-- -]ον.

6: εὐνομιῶτ[αν] *IC* in apparato; εὐνομιῶτ[-- -] *IC*.

Il collegio degli *eunomiotai*, attestato a Lato, Olous e forse Knossos con la denominazione di εὐνομία, a Polyrrenia con il nome di συνευνομιῶται (cf. *IC II 23 9*), anche ad Aptaera – come a Lato – si occupa dell’edificazione o della manutenzione di strutture pubbliche della città (cf. Guarducci 1933, pp. 201-205, Chaniotis 2008, pp. 114-116). Le competenze religiose che l’istituzione mostra di avere altrove – a Lato e Polyrrenia – suggerirebbero che ciò che è stato costruito o ristrutturato dagli *eunomiotai* sia uno o più edifici sacri, come sembra indicare l’espressione τᾶν λοιπᾶν πα[σᾶν], verosimilmente riferita all’oggetto di ἐπεμελήθηεν.

Quanto alla composizione del collegio, nella parte conservata dell’iscrizione è possibile identificare almeno cinque *eunomiotai*; il loro numero, tuttavia, potrebbe essere maggiore, similmente a quanto avviene a Lato, dove l’iscrizione completa *IC I 14 2* ne attesta nove, mentre il testo quasi completo di *IC I 16 21* ne ricorda sette.

Elementi istituzionali o altri termini rilevanti: *epimeletes* (*epimeleomai*), *eunomiotai*.

Figure 2. An inscription of the collection, *I.Cret. II 3 21*

dedicated to Lilian Jeffery’s work on Archaic Greek epichoric alphabets, *Poinikastas: Epigraphic Sources for Early Greek Writing* (<http://poinikastas.csad.ox.ac.uk>).

Istituzione	Termine attestato	Individuo	Tipologia	Ambito / Ruolo	Località	Periodo	Occorrenze
Damiorgos	δαμιοργέω	Eteon f. Archetos	Magistrato o funzionario	Dedicante	Aptera	E	seg_60_984.2
Damiorgos	δαμιοργός	A-	Collegio	Eponimo	Olous	E	seg_23_548.2
Damiorgos	δαμιοργός	Arsias	Magistrato o funzionario	Eponimo	Olous	E	[ic1_22_4.B.61]
Damiorgos	δαμιοργός	Autosthenes	Magistrato o funzionario	Eponimo	Olous	E	[ic1_22_4.B.1] ic1_22_4.B.19
Damiorgos	δαμιοργός	Botrynos	Collegio	Eponimo	Olous	E	[seg_23_549.1]
Damiorgos	δαμιοργός		Collegio	Eponimo	Kydonia	E	[seg_41_731.3]
Damiorgos	δαμιοργός		Collegio	Eponimo	Polyrrhenia	E	ic2_23_7.B.1
Damiorgos	δαμιοργός	Leukos	Magistrato o funzionario	Eponimo	Olous	E	ic1_22_4.A.31 ic1_22_4.A.35
Damiorgos	δαμιοργός	Onasandros f. Parmenon	Collegio	Eponimo	Polyrrhenia	E	ic2_23_7.A.1
Damiorgos	δαμιοργός	Sosos f. Tasskos	Collegio	Eponimo	Polyrrhenia	E	[ic2_23_8.1]

Figure 3. An excerpt from the index of institutional elements

4 Cretan Institutional Inscriptions at VeDPH

The Venice Centre for Digital and Public Humanities (VeDPH) was inaugurated in 2019 and belongs to the Department of Humanities of the Ca' Foscari University of Venice (UNIVE-DSU). The mission of the centre is the promotion of interdisciplinary methodologies for “the collaborative development of durable, reusable, shared resources for research and learning” (<https://www.unive.it/pag/39289>).

VeDPH not only promotes and funds new projects, but is also in charge of legacy projects developed at UNIVE-DSU over the past decades.

Since its foundation in 2019, the Venetian Detached Research Unit (URT) of the Institute for Computational Linguistics «A. Zampolli» of the National Research Council of Italy (CNR-ILC) has been working in collaboration with VeDPH within the *Archipelago DPH* project in order to ensure that the creation of new digital resources and the maintenance of the legacy ones are effectively durable, reusable and shared.

According to this vision, CLARIN-IT provides the necessary know-how through webinars and seminars at VeDPH, a state-of-the-art technological infrastructure to develop and test the new prototypes created by VeDPH affiliates, a suitable Web infrastructure for (permanently or temporarily) hosting the legacy projects developed at UNIVE-DSU and all the CLARIN tools and strategies to make new data and legacy data as FAIR⁵ as possible.

In this context, *Cretan Institutional Inscriptions* gave the opportunity to test this model of collaboration between VeDPH and CLARIN-IT through its Executing Institution, namely CNR-ILC.

⁵FAIR is an acronym for Findable, Accessible, Interoperable and Reusable (<https://www.go-fair.org>).

5 *Cretan Institutional Inscriptions* at CLARIN-IT

The Italian Consortium CLARIN-IT⁶ has a strong interest in the field of Digital Classics and aims at including a large part of resources for historical languages in its repositories (for an overview of the consortium at a whole see Nicolas et al., 2018). The Repository⁷ of the ILC4CLARIN Centre⁸ already contains important resources, such as the ALIM archive (Ferrarini, 2017), presented also at the CLARIN Conference 2020 (Boschetti et al., 2020), as well as many resources⁹ from the ERC project “LiLa: Linking Latin”¹⁰. The deposit and description of *Cretan Institutional Inscriptions* follow this path and increase the number of resources for Ancient Greek available in the CLARIN Virtual Language Observatory (VLO). Indeed, if VLO is queried for some of the main keywords used to describe *Cretan Institutional Inscriptions* in the ILC4CLARIN Repository (such as, for instance, *epigraphy* or *epigraphic*), only few resources are returned. The authors hope that *Cretan Institutional Inscriptions* pave the way for other similar initiatives to be described in Italy and other countries belonging to CLARIN-ERIC.

5.1 Organizational Aspects

In this section, the organization of the *Cretan Institutional Inscriptions* resources within the Italian node of CLARIN is described. The Web site dedicated to the dataset created within the Ph.D. project (<https://www.clarin-it.it/cretaninscriptions>)¹¹ is hosted by CLARIN-IT in the framework of the activities supporting projects and events in the Social Sciences and Humanities sector. The Web application to interact with the dataset (<https://ilc4clarin.ilc.cnr.it/cretaninscriptions>)¹² is offered as a service of ILC4CLARIN, the first CLARIN B-Centre of CLARIN-IT. The Persistent Identifiers (DSpace Handles) relating to the description sheets drawn up by the authors for the dataset¹³ and the Web application¹⁴ are available under a free license in the ILC4CLARIN Repository.

The strategy behind this organization is the following one: the *Cretan Institutional Inscriptions* collection has its GitHub repository (<https://github.com/IreneVagionakis/CretanInscriptions>), which contains the dataset, the software for the search engine, some customization, the licenses of use and the releases of the dataset. The authors decided to periodically deposit the various releases of the dataset in the ILC4CLARIN Repository as well, so that scholars can access the complete data without using the search engine. On the one hand, this approach guarantees the versioning of the dataset and the long term preservation of the data; on the other hand, it shares *Cretan Institutional Inscriptions* with the CLARIN community.

6 Dockerization

In the last decades a growing number of humanists exploited the digital ecosystem to improve their research and disseminate their results. Textual scholars learnt to build digital resources, to develop computational tools and to use research infrastructures, as well as to reuse generic frameworks, originally implemented for other domains of knowledge, but adapted to specific needs in Digital Humanities projects. Before the advent of the Docker system, two scenarios characterized the technological choice for new digital initiatives: 1) adopting resources and an infrastructure provided by some technological centre; 2) adopting proprietary resources and technologies within a single project. Both scenarios have advantages and disadvantages. As far as the first scenario (based on technological centres) is concerned, benefits lie mainly in soundness, availability, homogeneity, maintainability and security concerns about the

⁶<https://www.clarin-it.it>.

⁷<https://dspace-clarin-it.ilc.cnr.it/repository/xmlui>.

⁸<https://ilc4clarin.ilc.cnr.it/en>.

⁹The resources described are available in the Virtual Language Observatory (VLO): <https://vlo.clarin.eu/search?4&q=CIRCSE>.

¹⁰<https://lila-erc.eu>.

¹¹The PID for the URL is <http://hdl.handle.net/20.500.11752/1002>.

¹²The PID for the URL is <http://hdl.handle.net/20.500.11752/1003>.

¹³<http://hdl.handle.net/20.500.11752/OPEN-548>.

¹⁴<http://hdl.handle.net/20.500.11752/OPEN-550>.

technologies adopted. These advantages have the inevitable cost of experimenting new (if not immature) approaches, methods and tools. Actually, on the technological side, research activities were severely limited by these constraints. Indeed, the digital framework was often defined a priori and, therefore, hardly negotiable. For instance, relational databases were used alongside applications written in specific programming languages to build the Web Graphical User Interfaces (GUIs). As for the second scenario (proprietary technologies and resources), the flexibility offered in experimenting new methodologies, practices, models and resources often gave profitable outcomes. However, the innovative results were usually confined to a single initiative and almost never acknowledged within other similar initiatives. These last disadvantages were mainly due to the lack of well-defined standards, stable protocols, sufficient reliability and online availability of services.

A new scenario to build, share and run applications emerged recently. Indeed, Docker technology enables new DH applications to be packaged, published and deployed in single flexible units, called Docker images. The introduction of the Docker environment keeps the advantages offered by the aforementioned legacy scenarios but, at the same time, gets rid of their disadvantages. In fact, Dockerization provides a way to guarantee the availability of services, the maintainability of tools, the portability of applications, the flexibility of the environment, the scalability of the architecture and the consistency and reusability of the results of a digital project. In addition, the adoption of the Docker tool as the main deployment technology has two other advantages: 1) the DevOps methodology for agile development and continuous integration; 2) microservices for creating coherent components exposing data and services. Thanks to these properties, the typical problems of updating technical and data dependencies are successfully overcome, significantly improving data security.

Within this context, CLARIN-IT supported the *Cretan Institutional Inscriptions* team to adopt the DevOps methodology for the development, building and deployment of its digital resource (i.e. the collection of epigraphic texts encoded according XML/EpiDoc schema) as well as the Web application to interact with it (i.e. the EFES Web platform). This methodology increases *speed* of the development process of *Cretan Institutional Inscriptions* in each of its phases (text encoding, software implementation, testing and deployment on production servers). It also reduces *unexpected issues*, mainly due to different operative systems or different versions of software libraries. In addition, the DevOps methodology facilitates the *visualization* of the *Cretan Institutional Inscriptions* running components¹⁵ through a Web container manager. Thus, the CLARIN-IT team adopts Docker technology in order to implement a sound development workflow as well as long term preservation policies for the applications hosted, increasing their collaborative implementation and portability.

With the aim of maximizing the benefits of this methodology, CLARIN-IT runs the *Cretan Institutional Inscriptions* EFES application as a *stack of dockers containers*, managed through the Rancher environment, an open source Web container manager¹⁶. In this way, the different technologies and devices used (such as the Operative System, the Java Virtual Machine and the Web server) are separated from the technologies and devices of other applications running on the same server.

7 *Desiderata* for Future Developments

As with all work, although a lot has been done, a lot remains to be done. Among the main *desiderata* for the future there is undoubtedly the integration of data visualization systems through maps, relational graphs and timelines. This could be allowed by several existing open source tools and JavaScript libraries, among which Leaflet, Cytoscape.js and, above all, Palladio stand out.

Leaflet¹⁷, an open source JavaScript library for creating interactive maps, can be easily integrated inside EFES¹⁸, providing a filterable map view of the various institutional elements marked-up in the inscriptions.

¹⁵https://goto.docker.com/rs/929-FJL-178/images/20150731-wp_docker-3-ways-devops.pdf.

¹⁶<https://www.docker.com>, <https://rancher.com>.

¹⁷<https://leafletjs.com>.

¹⁸A precedent for this is the integration of Leaflet into the EFES-based site of *Fiscal Estate in Medieval Italy: Continuity and Change (9th-12th centuries)* project, *Fiscus*, <https://fiscus.unibo.it>.

Cytoscape.js¹⁹, an open source JavaScript library for creating interactive graphs, can also be easily integrated inside EFES²⁰, providing – through the usage of its force-directed fCoSE layout²¹ – highly dynamic relational networks.

Compared to the JavaScript libraries just mentioned, Palladio²² has the advantage of being an open source application developed specifically for the visualization of complex data resulting from historical research. Its visualization options include a Map view, a Graph view, a Table view and a Gallery view. In particular, its Map view is remarkably rich: in addition to the map itself, it includes the possibility of filtering data also according to a chronological criterion through the Timeline and Timespan filters. Such functionalities could allow to enhance the graphical performance of what can be obtained through a combination of multiple EFES search filters, which already allow a cross-search of geographical and temporal data and information relating to the institutions mentioned (as well as other variables relating to the epigraphic sources and their contents). The following figures (Figs. 4-8) offer some examples of what types of visualizations can be obtained through the Palladio Web application, using respectively its Map view (Fig. 4), its Map view with a Timespan filter (Figs. 5-6) and its Graph view (Figs. 7-8) applied to some case studies relating to the institutions attested as decreers and dedicants and to the full picture of the attestations of tribes and *agelai*.

In addition to the Map, Graph and Timeline views, in the future it would be useful to develop some other features, such as the implementation of an API interface allowing an export of the data also in RDF²³ or JSON formats and, from the point of view of contents, the inclusion of the English translations of the inscriptions.



Figure 4. Geographical distribution of the attestations respectively of *kosmoi* (top left), *boule* (top right), *demos* (bottom right), and *polis* (bottom left) as decreers (visualization on Palladio - Map view)

¹⁹<https://js.cytoscape.org>.

²⁰A precedent for this, again, is the integration of Cytoscape.js into the EFES-based site of *Fiscal Estate in Medieval Italy: Continuity and Change (9th-12th centuries)* project, *Fiscus*, <https://fiscus.unibo.it>.

²¹<https://github.com/iVis-at-Bilkent/cytoscape.js-fcose>.

²²*Palladio. Visualize complex historical data with ease*, <http://hdlab.stanford.edu/palladio>, <https://github.com/humanitiesplusdesign/palladio-app>.

²³Kiln and EFES already provide some basic functionalities for handling RDF data: see <https://kiln.readthedocs.io/en/latest/tutorial.html#querying-rdf> and <https://kiln.readthedocs.io/en/latest/rdf.html>.

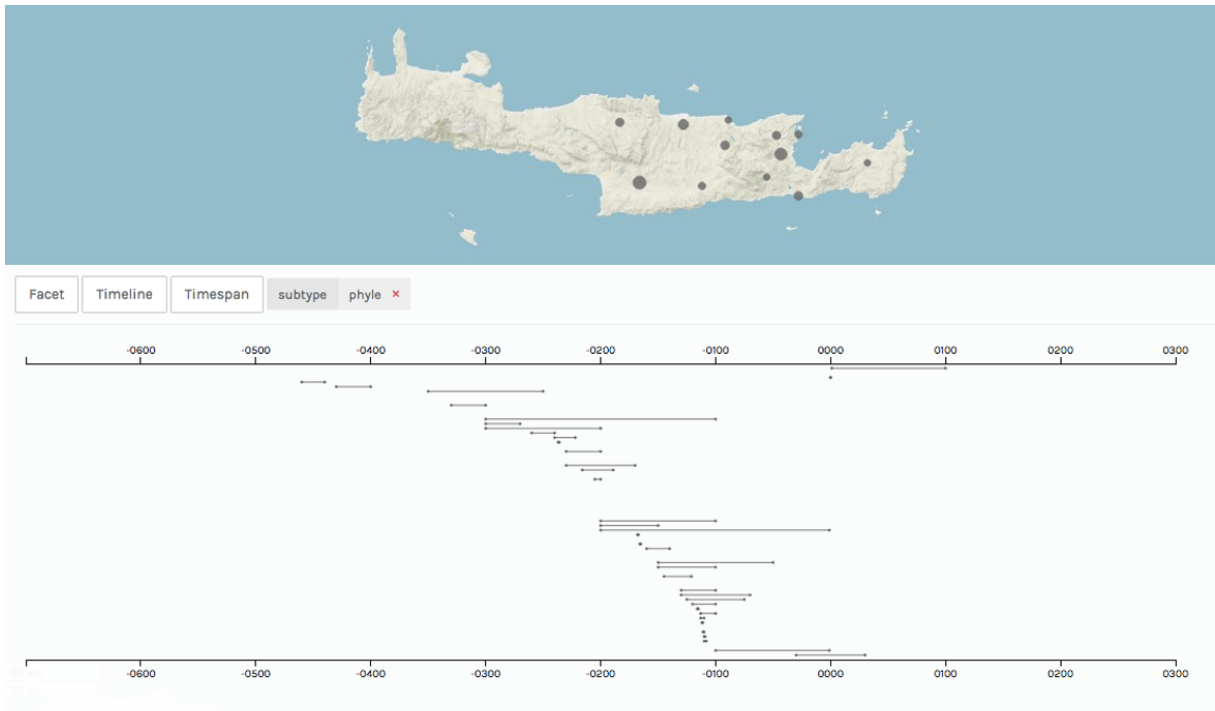


Figure 5. Geographical and chronological distribution of the attestations of tribes (visualization on Palladio - Map view including a Timespan filter)



Figure 6. Geographical and chronological distribution of the attestations of *agelai* (visualization on Palladio - Map view including a Timespan filter)

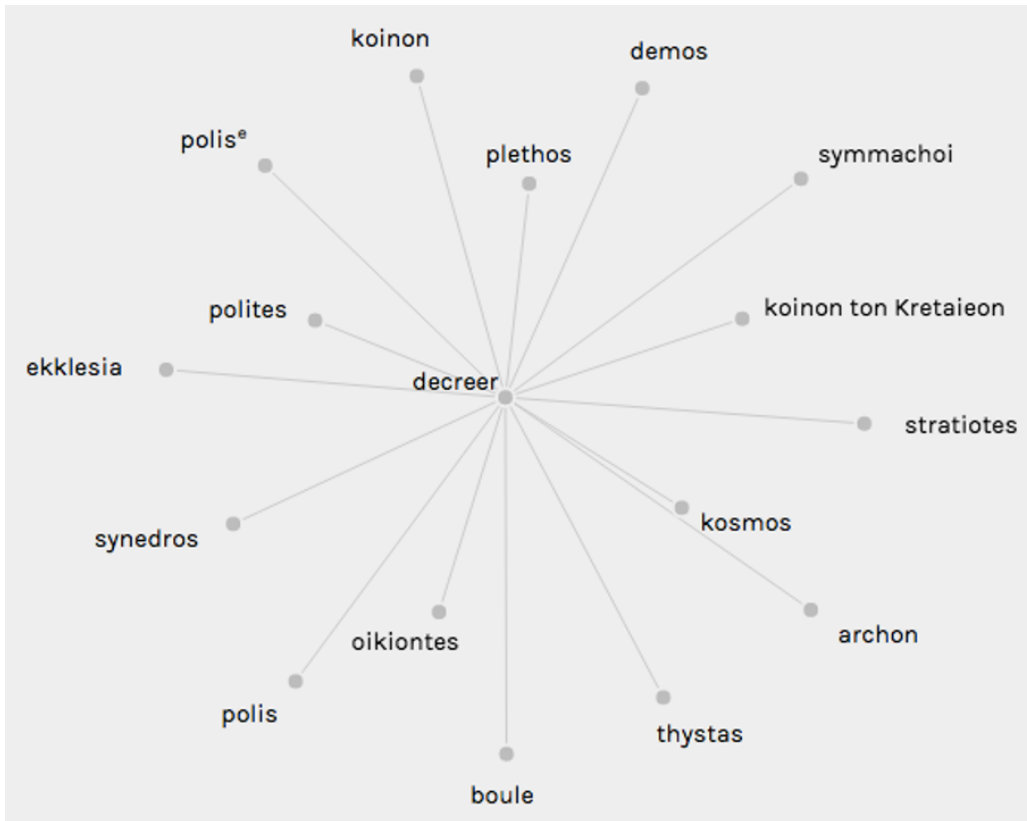


Figure 7. Institutions attested as decrees (visualization on Palladio - Graph view)

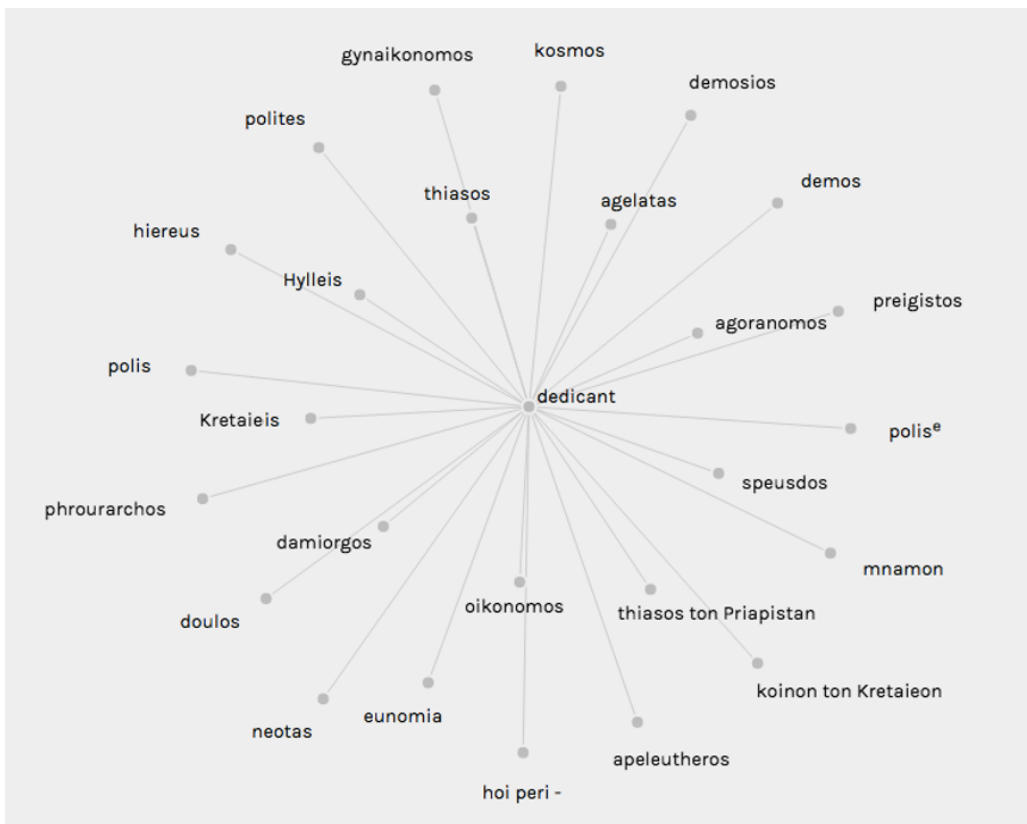


Figure 8. Institutions attested as dedicants (visualization on Palladio - Graph view)

8 Digital Epigraphy and the CLARIN Community: The Role of K-Centres

The original focus of CLARIN is on language resources, such as dictionaries and treebanks, and textual resources, such as literary corpora, used in particular as a test to study linguistic phenomena in a context. Recently, however, the interest of the CLARIN community in the domain of Digital Humanities significantly increased, as it emerged in the last CLARIN Annual Conferences, in CLARIN initiatives such as the CLARIN Café series and in initiatives supported by CLARIN such as the last Annual Conference of the Italian Association of Digital Humanities and Digital Culture (AIUCD)²⁴.

By widening its horizons, the CLARIN community has enlarged its interdisciplinary competence towards the digital representation of cultural artifacts, such as ancient and modern epigraphs, Medieval manuscripts, ethnographic oral archives etc. The processing of fragmentary texts, Information Retrieval applied to the text of scholarly editions with variant readings in the critical apparatus and the mapping of linguistic annotations on facsimile of historical documents are just a few examples of the new challenges that the CLARIN Infrastructure has to face.

On the one hand, CLARIN helps the disciplinary communities of digital philologists, digital epigraphers and digital historians to share best practices relating to digitalization processes in Digital Humanities. In particular, these best practices concern the digitization of images, the transcription of primary sources, the encoding of texts, the annotation of linguistic, stylistic and philological phenomena, the visualization of scholarly editions, the creation of search engines and the publication and reuse of literary and documentary data. Indeed, the use of language resources is transversal to different disciplines. New epigraphic corpora enrich the repertoires of attested inflected forms, which feed up the spellcheckers used to improve the OCR post-processing applied to corpora of secondary sources etc. On the other hand, CLARIN has the opportunity to extend the infrastructure in order to meet the specific needs of the humanists, in particular the interrelation between text and document, the non-sequentiality of the variorum edition and the diachronic perspective.

CLARIN Knowledge Centres²⁵ are suitable instruments to share interdisciplinary competence that, in order to accomplish each phase of a digital project, has to be coordinated. Among them, the new Knowledge Centre for Digital and Public Textual Scholarship (DiPText-KC)²⁶, which is the second CLARIN K-Centre of CLARIN-IT, is specifically devoted to the scientific representation of literary and documentary texts as well as to their elaboration and publication.

Cretan Institutional Inscriptions gave us the opportunity to test the organization of DiPText-KC in order to put the project leader in close contact with a team of experts in the DevOps technologies distributed among Pisa, Venice and other European DH Centres. Indeed, Digital Epigraphy is an interesting test bench for verifying the cross-fertilization between language infrastructures and Digital Humanities, since, in order to study the historical context etc., it involves the description of the material support, the study of (possibly fragmentary) texts and references to the secondary literature.

Being a solid project based on EpiDoc, *Cretan Institutional Inscriptions* adopts standard protocols for encoding its textual resources. DiPText-KC addressed the project leader to publish the software source code and textual data separately and under open licenses as well as to make data and metadata compliant to the FAIR principles in order to maximize their reusability.

9 Conclusion

With this paper the authors aimed to highlight how CLARIN-IT is opening up to areas of Digital Humanities that, until few years ago, were not central to the CLARIN world. Indeed, CLARIN has always demonstrated interest not only in the living language but also in literary texts, as it is evidenced by the high number of corpora in the CLARIN repositories. However, its interest in Ancient Greek and Latin Digital Epigraphy and Papyrology is more recent. A search in VLO with the keyword “epigraph* Greek OR Latin” or with the keyword “papyr* Greek OR Latin” shows that most of the records is very recent (around 2020). These disciplines give CLARIN the possibility to reason about new use cases to extend

²⁴<https://aiucd2021.labcd.unipi.it/en/home-english>.

²⁵<https://www.clarin.eu/content/knowledge-centres>.

²⁶<https://diptext-kc.clarin-it.it>.

the metadata set with information needed by epigraphers and papyrologists, such as the geolocalization of the document, the presence or not of a facsimile in the collection, the literary genre to which texts belong, if a text is in poetry or prose etc.

Federated Content Search can also benefit from an expansion to new types of documents, especially in the way of citing the occurrences returned by a query and in the way of dealing with fragmentary texts.

The authors hope that a project such as *Cretan Institutional Inscriptions*, which contain both a critical edition and a specific visualization tool, can contribute to widen the bridge between purely linguistic interests and other areas of the Humanities inside the CLARIN world, where this important connection is often missing. It is to this end, in fact, that the authors are implementing a DevOps methodology and a Docker infrastructure aimed at hosting such a kind of initiatives.

10 Acknowledgment

The authors are grateful to Alessandro Enea (CNR-ILC) for the technical support to the porting of *Cretan Institutional Inscriptions* on the CLARIN-IT servers. The authors also wish to thank Franz Fischer (VeDPH) for the organizational support and Gabriel Bodard and Pietro Liuzzo for their helpful feedback about EFES customization and data visualization.

References

- Bodard, G. and Yordanova, P. 2020. Publication, Testing and Visualization with EFES: A Tool for All Stages of the EpiDoc XML Editing Process. *Studia Universitatis Babeş-Bolyai Digitalia*, 65(1):17–35, Dec.
- Boschetti, F., Del Gratta, R., Monachini, M., Buzzoni, M., Monella, P., and Rosselli Del Turco, R. 2020. “Tea for Two”: The Archive of the Italian Latinity of the Middle Ages Meets the CLARIN Infrastructure. In *CLARIN Annual Conference 2020*, pages 121–125. CLARIN-Virtual Edition.
- Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., and Vanderbilt, S. e. a. 2020. EpiDoc Guidelines: Ancient Documents in TEI XML (Version 9.2).
- Ferrarini, E. 2017. Alim ieri e oggi. *Umanistica Digitale*, 1(1).
- Liddell, H., Scott, R., and Jones, H. S. 1925-1940. *A Greek-English Lexicon. Ninth Edition*. Oxford Clarendon Press.
- Liddell, H., Scott, R., and Jones, H. S. 2011. *The Online Liddell-Scott-Jones Greek-English Lexicon*. Thesaurus Linguae Graecae.
- Nicolas, L., König, A., Monachini, M., Del Gratta, R., Calamai, S., Abel, A., Enea, A., Biliotti, F., Quochi, V., and Stella, F. 2018. CLARIN-IT: State of Affairs, Challenges and Opportunities. In *Selected Papers from the CLARIN Annual Conference 2017*.
- Vagionakis, I. 2021. Cretan Institutional Inscriptions. A New EpiDoc Database. *Journal of the Text Encoding Initiative*.

Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts

Elena Volodina, David Alfter
University of Gothenburg, Sweden
name.surname@gu.se

**Therese Lindström Tiedemann,
Maisa Lauriala, Daniela Piipponen**
University of Helsinki, Finland
name.surname@helsinki.fi

Abstract

We present the results of a manual evaluation of the performance of automatic linguistic annotation on three different datasets: (1) texts written by native speakers, (2) essays written by second language (L2) learners of Swedish in the original form and (3) the normalized versions of learner-written essays. The focus of the evaluation is on lemmatization, POS-tagging, word sense disambiguation, multi-word detection and dependency annotation. Two annotators manually went through the automatic annotation on a subset of the datasets and marked up all deviations based on their expert judgments and the guidelines provided. We report Inter-Annotator Agreement between the two annotators¹ and accuracy for the linguistic annotation quality for the three datasets, by levels and linguistic features.

1 Introduction

In the current project, *Development of grammatical and lexical competences in immigrant Swedish*,² we explore profiling of lexical and grammatical competences among second language (L2) learners of Swedish based on two corpora. The coursebook corpus, COCTAILL (Volodina et al., 2014), and the L2 Swedish learner corpus, SweLL-pilot (Volodina et al., 2016), are used for qualitative and quantitative analysis of lexical and grammatical categories that L2 learners are exposed to or produce themselves. The texts in the two corpora have been automatically annotated with linguistic information using the Sparv-pipeline (Borin et al., 2016) which is an essential part of the CLARIN infrastructure for the Swedish language. Sparv, in turn, relies on the gold annotation standards from the Stockholm Umeå Corpus (SUC) (Ejerhed et al., 1997) and on the theoretical framework in the Saldo lexicon (Borin et al., 2013). Since the process of linguistic annotation is performed automatically, we need to evaluate to which degree we can expect the results of the annotation to be reliable, so that our theoretical generalizations and conclusions about language learning can factor that in. For this reason, we performed a manual “annotation quality check” of Part-of-Speech (POS) tagging, lemmatization, dependency annotation, identification of multi-word expressions (MWE) and word sense disambiguation (WSD) which we report in this paper.

Previous work suggests that performance of automatic pipelines trained on native language models is non-optimal on L2 language due to a large number of non-words, deviating syntactic patterns and statistical distributions in L2 production (Štindlová et al., 2012). Rubin (2021) shows that the performance of two independent parsers for Dutch drops by $\approx 7\text{--}8\%$ on L2 learner data compared to first language (L1) data. Krivanek and Meurers (2013) have similar results for L2 German, with $\approx 6\%$ drop in LAS (labeled attachment scores) for dependency parsing of L2 German. Ott and Ziai (2010) have observed that not all L2 deviations have an equally drastic impact on automatic linguistic annotation, e.g. deviations in morphology and word order do not influence the accuracy of POS tagging or syntactic parsing, whereas omission of syntactically important relations, such as subjects and verbs, yields incorrect parses. Meurers and Wunsch (2010) discuss the need for theoretical analysis of linguistic features in learner language with implications for automatic L2 annotation. For example, the three criteria for assigning a part

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

¹Note, though, that dependency annotation was checked by one annotator only

²Riksbankens jubileumsfond P17-0716:1, project homepage: <<https://spraakbanken.gu.se/en/projects/l2profiles>>

of speech (POS) - lexico-semantic, morphological and distributional - are not always applicable to the L2 data, e.g. I was **choiced* for a job; He walked **rapid*; where one or more of the criteria are not followed. However, there is no common consensus (and very little discussion) about which automatic fallback strategies should be preferred in case of automatic annotation of non-native language or whether the principles of L2 annotation should be more drastically revised (Meurers and Wunsch, 2010).

A very dangerous trap in annotation of learner language is to start encoding what the learner meant (which is subjective in nature) rather than objectively describing what has been used. To ensure objectivity in L2 POS-tagging, it might be best if all the three criteria could be encoded separately. This would mean that for the word **rapid* in *He walked *rapid*, three POS codes could be assigned: lexical-POS: adjective; morphological-POS: adjective; and distributional-POS: adverb.

To our knowledge, no evaluation of automatic linguistic annotation on Swedish L2 data has been done yet. In the present paper we present the results of such an evaluation for the Sparv pipeline and our conclusions regarding the applicability of the Sparv pipeline for analysis of L2 data. This experiment complements and extends several investigations of the Sparv pipeline where Sparv has been analyzed from the point of view of automatic tools, models and modules (Ljunglöf et al., 2019), and its performance has been *automatically* evaluated in relation to *native* language (L1) varieties (Berdicevskis, 2020a; Berdicevskis, 2020b), whereas we examine the reliability of annotations *manually* and on several types of language – L1, L2 original and L2 normalized (i.e. corrected). We analyze the performance of the tool by categories and subcategories, as well as in relation to different L2 proficiency levels.

Despite Swedish being the focus of this experiment, we expect our findings to be generalizable to other languages and to the performance of other pipelines on non-native language samples. It is an important study for CLARIN since it evaluates how well part of the CLARIN infrastructure works for both L1 and L2 Swedish, thereby assessing the need for improvements to the current pipeline for Swedish.

2 Notes on Linguistic Terminology

Notion of Lexical Items The way researchers operationalize the construct of a “word” influences the way word statistics and frequency counts are collected and the way different aspects of individual items are analyzed. This has a direct impact upon the application of the collected statistics (Gardner, 2007). One of the most common ways to work with words is based on *lemmas* (=base forms of a word, e.g. *file*) and its derivative version *lemgrams* (=base form + POS, e.g. *file*, *verb*). There are different ways to define the notion of lemgrams. In our case we rely on the operationalization of lemgram in the Saldo lexicon (Borin et al., 2013) which is used in the Sparv-pipeline (Borin et al., 2016, p.1): “A lemgram is a lexical identifier which refers to an inflection table in the SALDO lexicon (Borin et al., 2013), which provides linkages between lemgrams and word sense identifiers, although the relation is many-to-many.” This means that Sparv can differentiate between words of the same part of speech if they belong to different inflectional paradigms, e.g. between the verb *hang*-*hanged* and the verb *hang*-*hung*; however, if both the base form **and** the inflectional paradigm are shared, homographical items are not automatically differentiated, e.g. *fil* ‘file on a computer’ vs *fil* ‘driving lane’. Sparv therefore provides a pointer to several possible senses of each identified lexical item (e.g. to different senses of all possible “file” nouns with the same inflectional paradigm). Even word senses are derived from the Saldo lexicon, with regards to their identifiers, descriptors and number of senses per lemgram, and are used in the module for *word sense disambiguation* in the Sparv pipeline.

Notion of a Single-Word Lexical Item *Lemgram* is usually understood as a set of word forms having the same base form and belonging to the same POS, e.g. all occurrences of the word forms *flicka*, *flickas*, *flickan*, etc. are counted together since they have the same base form *flicka* ‘girl’ and the same part-of-speech *noun*. The Sparv annotation takes this a step further, where lemgrams are also differentiated based on inflectional paradigms encoded in Saldo, so that *val* (noun, -et; the neuter gender, 6th declension; ‘election; choice’) and *val* (noun, -en, -ar; the uter gender, 2nd declension; ‘whale’) count as two different items in frequency statistics. Besides, due to the recent development in word sense disambiguation approaches for Swedish (Nieto Piña, 2019), it is now possible to collect triples of identifying

information for each lexical item, namely lemma+POS+sense. Thus, for the lexical item *gräva*, verb, we are able to collect frequencies separately for the sense *dig a hole* and for the sense *do research*.

Notion of a Multi-Word Expression The concept of Multi-word expression (MWE) is both broad and vaguely defined. The literature abounds in different terms with similar meanings: *collocations* (Bhalla and Klimcikova, 2019), *phraseological units* (Paquot, 2019), *lexicalized phrases* (Sag et al., 2002), *formulaic sequences* (Wray, 2005), etc. The definition of multi-word expressions in our study is inherited from the Saldo lexicon which is used in the Sparv annotation pipeline, where Saldo forms the lexical knowledge-base. The Saldo definition of MWEs is based on semantic-orthographic principles, i.e. an MWE consists of two or more orthographically defined lexical items, while exhibiting a certain (varying) extent of semantic non-compositionality (Borin, 2021). Each MWE is a lemmagram of its own, can have several senses and falls into one of the three structurally-defined broad categories: contiguous, non-contiguous or constructions (Borin, 2021, p.223). However, constructions, which by definition contain open placeholder e.g. *på X bekostnad* ‘on X’s account’, are not yet fully integrated into Saldo, and are therefore not yet automatically processed by the Sparv pipeline either. We accept the Saldo definition of MWEs at face value for this particular investigation limiting ourselves to the first two types of MWEs (see, however, Alfter et al. (2021) for our more refined taxonomy developed within the context of the current project based on the first two MWE types in Saldo).

Part-of-Speech Categories As is clear from the descriptions above, we are focusing on the analysis of annotation tags (and their interpretation) present in the Sparv annotation output, even though they do not always reflect the way we may want to define the categories which they represent. The same concerns part-of-speech (POS) categories.

There are two POS taxonomies used in the output of Sparv: one coming from the model trained on SUC (Gustafson-Capková and Hartmann, 2006), a gold-annotated corpus, with 22 POS categories³; and the other based on the Saldo lexicon (Borin et al., 2013), with 37 POS categories⁴. The analysis in this experiment is focused on the SUC-based POS tags (see Appendix A for an overview). There is an option to convert SUC-based POS tags into the universal tagset⁵ (Petrov et al., 2011), but the conversion is not fully reliable. Not all POS categories used in the Sparv output correspond to the part-of-speech defined in the Swedish Academy Grammar (SAG) (Teleman et al., 1999), which is the most authoritative description of Swedish grammar. The difference is especially notable in relation to *determiners* which are used in the SUC tagset, but are not among POS categories in SAG. Another difference concerns adverbial usage of neuter adjectives (e.g. *högt*) which in SUC are treated as adverbs but as adjectives in neuter form in SAG (i.e. adjective *hög* + neuter inflection *-t*). The conflicting theoretical views on POS categories may have prompted unnecessary corrections by the annotators.



Figure 1: Syntactic tree based on Sparv annotation.

Dependency Relations Categories used by Sparv come from the MAMBA tagset⁶ used in the Swedish treebank Talbanken (Nivre et al., 2008). The Mamba tagset contains sixty-five (65) tags including fourteen (14) tags describing punctuation (see Appendix B for a full taxonomy). The dependency relations (DepRels) are split into Root (or head) and Relations (or syntactic functions), e.g. *subject*, *finite verb*,

³<https://spraakbanken.gu.se/korp/markup/msdtags.html>

⁴<https://spraakbanken.gu.se/en/resources/saldo/tagset>

⁵<https://universaldependencies.org/u/feat/index.html>

⁶https://cl.lingfil.uu.se/~nivre/swedish_treebank/dep.html

direct object, agent. No conversion to the Universal Dependency Relations⁷ (De Marneffe et al., 2014) is offered by Sparv. DepRel tags are used to build syntactic trees (see Figure 1) where syntactic relations are shown through arrows, while POS tags are shown in squares.

3 Experiment Setup

Figure 2 shows the main steps in the experimental setup. We started with three main hypotheses (subsection 3.1), selected three datasets appropriate for testing our hypotheses (Section 3.2), processed all the datasets with the Sparv pipeline (Section 3.3), and manually checked the automatic annotation (Section 3.4). The choice of evaluation metrics and quantitative analysis of the results are given in Section 4, followed by a qualitative analysis in Section 5.

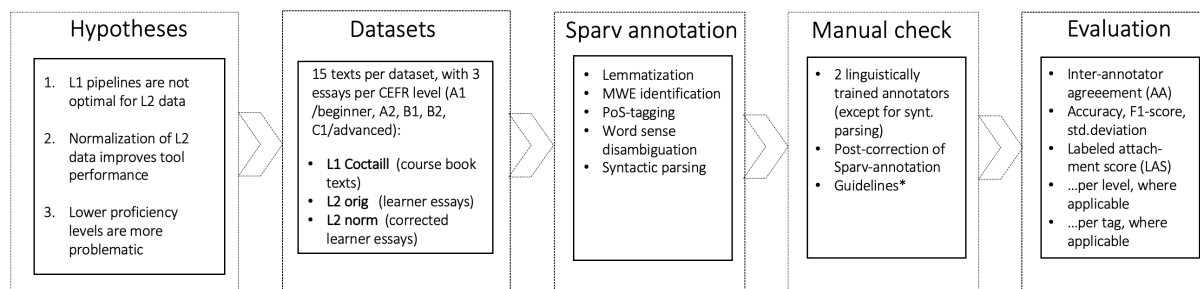


Figure 2: Overview of the experiment setup.

(*)Guidelines: <https://tinyurl.com/bdhsukys>

3.1 Hypotheses

As is obvious from the short description of the automatic linguistic annotation of learner language given in the Introduction, there is a need to explore the reliability of automatic pipelines further, to assess the needs to adapt the pipelines for L2, and to discuss the implications of the results for L2 theoretical studies and practical applications. Our hypotheses for this experiment are:

- Pipelines trained on a standard language (L1) do not perform as well on non-standard language varieties such as learner language (e.g. L2 learner production).
- Normalization of non-standard language, e.g. through error correction, improves tool performance.
- The need for normalization (cf "correction") is especially critical for L2 texts written by learners at lower proficiency levels since they are likely to contain a higher level of misspellings, wrong words and syntactic discrepancies in comparison to the standard.

Even though some of the claims above appeal to common sense, they need to be confirmed explicitly and there is a need for an estimate of how well, or how poorly, the automatic annotation works in order to know how it can be reused for research, CALL and other scenarios.

3.2 Datasets

To address the hypotheses above, we selected 15 texts per language variety which we are interested in – namely, native language used in L2 Swedish course books (*L1 Coctail*), L2 essays (*L2 orig*) and corrected L2 essays (*L2 norm*) – so that they represent five levels of proficiency with three texts per level for each dataset. The levels are defined in accordance with CEFR, the Common European Framework of Reference (Council of Europe, 2001), in our datasets covering five of the six levels: A1 (beginner), A2, B1, B2 and C1 (advanced). C2 was excluded due to a lack of data in the source corpora.

Care was taken to select texts of different genres and topics to avoid biases. Only texts containing at least one MWE according to the Sparv annotation were selected. Learner essays in the *L2 orig* dataset

⁷<https://universaldependencies.org/u/dep/index.html>

represent speakers of different native languages – namely: Chinese, English, Finnish, Flemish, Lithuanian, Macedonian, Persian, Romanian, Serbian, Somali, Spanish, Tigrinya, Vietnamese – to avoid potential influence of L1 on L2 usage. Detailed statistics over the datasets are available in Appendix C.

L1 Coctail *L1 Coctail* is a dataset representing native language and contains 2190 tokens (incl. punctuation) per 15 texts. These texts comprise various genres (narrations, facts, evaluation, dialogues, letters, poems) and different topical domains (traveling, languages, culture and traditions, relations with other people, etc.). The dataset is based on *COCTAILL* – a corpus of course books (Volodina et al., 2014), where each chapter has been marked with the level of proficiency at which it could be used in the teaching of L2 Swedish. The CEFR levels are represented by three texts per level.

L2 orig *L2 orig* is a dataset that contains 4012 tokens (incl. punctuation) per 15 essays, with three essays per CEFR level, covering several genres (narration, evaluation, argumentation, etc.) and topical domains (personal identification, daily life, travel, house and home, culture and traditions, etc.). *L2 orig* is a subset of the *SweLL-pilot* – a corpus of learner-written essays (Volodina et al., 2016) collected from three different schools/test bodies, and also marked with CEFR levels.

L2 norm *L2 norm* is a dataset containing 3955 tokens (incl. punctuation) per 15 essays and consists of the same essays (or in two cases of comparable essays, e.g. of the same topic and genre) as in *L2 orig*, but normalized for errors and deviations to reflect the current norms of the target language. The normalization was performed using the SVALA tool (Wirén et al., 2019), by a linguistically trained L1 Swedish speaker following the normalization guidelines from the SweLL-project (Rudebeck et al., 2021).

3.3 Sparv Pipeline

The Sparv pipeline⁸ (Borin et al., 2016), consists of several modules, sequentially applied to the Swedish data input. In version 3.0, analyzed by us, for *lemmatization*, the Saldo lexicon (Borin et al., 2013) returns lemgrams including potential MWEs and a list of associated senses. *Senses* are disambiguated using an algorithm developed by Nieto Piña (2019) based on Saldo senses. For *POS tagging*, Sparv uses HunPos (Halácsy et al., 2007) trained on the SUC 3.0 corpus (Ejerhed et al., 1997). For *syntactic annotation*, the MaltParser (Nivre et al., 2007) is used, trained on the Swedish Talbanken (Nilsson et al., 2005).

A new Sparv version (4.0) was released for public use in 2021,⁹ where the POS tagger and syntactic annotation are changed to Stanza (Qi et al., 2020) with new models. According to Berdicevskis (2020a) and Berdicevskis (2020b), annotation for syntactic relations and POS tagging in versions 4.0 and above should have a higher accuracy than previous versions. The newer versions of Sparv continue using models trained on SUC and Talbanken, which means that the tagsets for both POS and DepRels are still the same.

3.4 Manual Check

Two linguistically trained assistants, one an L1 Swedish speaker and one an advanced L2 Swedish speaker (L1 Finnish), manually analyzed the automatic tags of the three datasets, introducing corrections where necessary. Assistants were equipped with guidelines¹⁰ and were in regular contact with one of the researchers for discussions, which cleared up uncertainties and led to clarifications in the guidelines. They performed the check using separate spreadsheet files to avoid influencing each other. Instructions were specific for each linguistic feature.

The rule of thumb for the annotation check was to start from a *positive assumption* that the Sparv-pipeline’s suggestions are correct, and introduce corrections only if necessary and motivated. With regards to annotation of learner essays, it meant *disregarding* the perspective of “what the learner meant” and assessing the output of the pipeline from a formal point of view, i.e. what it had been fed.

Most problems arose from the conceptual interpretation of the task in relation to the *L2 orig* dataset, namely, what to consider correct or incorrect output from the pipeline. Consider the following example:

⁸History of Sparv-releases: <https://github.com/spraakbanken/sparv-pipeline/releases>

⁹<https://github.com/spraakbanken/sparv-pipeline/releases/tag/v4.0.0>

¹⁰<https://docs.google.com/document/d/1W9gcwRwFJ7-DsAC6cf6BHUoEivt73r-XWCV1oKS6xV8/edit?ts=5f3518d7#>

[1] Jag tycker om spela fotbol , sinima , cykler och TV-speL . (A1 level)
 ‘I like to play fotbal , ?swinim / ?sinima , bykes¹¹ and TV-gameS .’ (translation tries to replicate the errors in the original variant. *sinima* may be an attempt to write *simma* ‘to swim’, but since the learner lists hobbies it could also be an attempt to write the English word ‘cinema’ in a more Swedish way instead of the corresponding Swedish word *bio*.)

The word *cykler*¹² ‘bikes’ could be interpreted as either the present tense of the verb “to bike”¹³, *cyklar*, or the plural form of the noun “a bike”, *cyklar*. The use of TV-speL ‘TV-gameS’ suggests that a noun is a possible alternative in a list together with the noun TV-speL. However, a verb is also a fully legitimate alternative, as a part of a list together with the verb *spela* (*fotbal), ‘play (football)’. The assistants have annotated this output differently – one correcting the Sparv-suggested *noun*-tag for *cykler* with a *verb*-tag, the other accepting the *noun*-suggestion as the right one. Similarly, the misspelled word *sinima* was automatically tagged as a noun, and accepted as such by one of the assistants, but changed to a verb by the other. These examples show the problems of dealing with learner data and potential reasons for disagreements between the annotators. Both interpretations above are equally possible and equally close to the original.

The example below is easier to interpret and does not cause disagreement between the annotators. One learner produced a misspelling of the preposition *enligt* ‘according to’ which was tagged as an *adjective*, most probably due to its morphological form in conjunction with the position in the sentence:

[2] Enligt ungmmedia.se, är... (C1 level)
 ‘Accordin to ungmmedia.se, is...’ (translation tries to preserve the errors in the original variant.)

Both annotators corrected Sparv-suggested tag *adjective* to *preposition*. In standard Swedish, the first position of a sentence is most likely to contain a subject, often consisting of a noun phrase which can contain an adjective. However, to a human annotator, the similarity of *enligt* to the word *enligt* was obvious; it was also obvious that there is no adjective that is similar to this. This motivated the correction to the pipeline’s output and suggests a need to check for lexical similarity in POS-tagging.

4 Results

On completion of the check, we analyzed the number of deviations discovered during the manual check and inspected their nature per linguistic category in each dataset and in relation to the proficiency level and tagset, where appropriate. Below, we report these results using precision, F1-score and LAS measures (averaged over the two annotators for all tasks except syntactic parsing/DepRel annotation which was checked by only one annotator). For word sense disambiguation, we have additionally computed a baseline using the first sense in all cases.

Inter-Annotator Agreement To put the reported results into perspective, we calculated inter-annotator agreement (IAA) for the two annotators using Krippendorff’s alpha (Krippendorff, 2004) for MWEs, and pairwise agreement for Lemma, POS and Sense, see Table 1. Pairwise agreement is calculated on a token basis, and we count only whether a change has been made to the original annotation or not.

Corpus	Lemma	POS	Sense	MWE
L1 Coctail	0.95	0.97	0.88	0.85
L2 orig	0.94	0.95	0.88	0.74
L2 norm	0.95	0.97	0.90	0.89

Table 1: Pairwise agreement for Lemma, POS and Sense; Krippendorff’s alpha for MWE

The agreement lies over 0.8 for most of the datasets and denotes high agreement. We see that values for *L2 orig* is nearly always lower than for the other datasets; reasons for that have been briefly touched

¹¹since the original *cykler* is a misspelling, we mock a misspelling in the English version of the word *bike*

¹²Note that *cykler* is a misspelling, too.

¹³While “to cycle” might be a more idiomatic translation, we want to illustrate the homonymy between word classes here

upon in Section 3.4. Most disagreements appear in the evaluation of the MWE identification, with the lowest at 0.74 for *L2 orig*. The intersection of corrections introduced by both annotators is high. Still we see that one annotator is better at noticing grammatical MWEs (e.g. *trots att* ‘even though’ and the other is better at spotting light verb constructions (e.g. *få barn* ‘have a child/children’) and this causes disagreement, but enriches the results of the check.

Corpus	Lemma	POS	DepRel
L1 Coctail	0.93 (0.0)	0.98 (0.0)	74.49
A1	0.96 (0.02)	0.98 (0.0)	75.93
A2	0.98 (0.03)	0.97 (0.02)	72.51
B1	0.94 (0.0)	0.97 (0.0)	76.65
B2	0.89 (0.0)	0.97 (0.01)	71.05
C1	0.92 (0.01)	0.97 (0.0)	76.31
L2 orig	0.90 (0.02)	0.95 (0.0)	63.01
A1	0.89 (0.01)	0.92 (0.0)	51.66
A2	0.89 (0.0)	0.94 (0.0)	57.18
B1	0.91 (0.02)	0.96 (0.01)	60.42
B2	0.92 (0.04)	0.97 (0.01)	67.53
C1	0.92 (0.03)	0.97 (0.01)	69.18
L2 norm	0.93 (0.02)	0.97 (0.0)	69.02
A1	0.95 (0.0)	0.98 (0.01)	67.23
A2	0.92 (0.0)	0.96 (0.0)	69.30
B1	0.95 (0.02)	0.98 (0.01)	70.53
B2	0.92 (0.03)	0.98 (0.01)	71.52
C1	0.92 (0.02)	0.97 (0.0)	66.80

Table 2: Lemmatization and POS tagging: precision and standard deviation; Dependency: LAS

4.1 Automatic Lemmatization, POS-tagging and Dependency Annotation

Table 2 summarizes the results regarding the quality of the automatic annotation (i.e. how often the two annotators corrected automatically assigned tags) for lemmatization, POS tagging and Dependency Relations. Lemmatization and POS-tagging are evaluated in terms of precision (number of correct items by total number of items), averaged over the two annotators. Dependency annotation is evaluated using micro-averaged (i.e. token-based) Labeled Attachment Score (LAS) (Kübler et al., 2009).¹⁴

Automatic Lemmatization Results for automatic lemmatization show that it is very successful, with 93% precision on average for *L1 Coctail*. As expected, the number decreases in *L2 orig* in comparison to *L1 Coctail*, resulting in 90% precision; and after normalization it increases to 93% in *L2 norm*, the same level as in *L1 Coctail*. We also see the expected tendency of quality increase in the *L2 orig* by proficiency level. As learners become more proficient they write in a way that can be expected to be closer to L1, a language containing less discrepancies and hence easier to annotate automatically with tools trained on L1 data. The fact that we do not see the same increase in *L1 Coctail* is probably due to the fact that language presented as reading materials to learners at more advanced levels can contain more specialized vocabulary, some of which might not be in Saldo. It is interesting that similarly we also do not see an increase over all levels in the *L2 norm*. But this correlates with the L1 data and it is notable that C1-level in this data is as well lemmatized as the L1 data.

Part-of-Speech Tagging Results for POS-tagging are systematically high across all datasets, with the average top 98% for *L1 Coctail* and the lowest average result of 95% in *L2 orig*. Normalization of learner

¹⁴Note that the dependency annotation was checked by one assistant, while the rest of the annotation was checked by two.

essays improves the results for POS by 2 points. Just like for lemmatization there is a clear improvement in the POS-tagging on higher levels in the *L2 orig*, which reaches as high precision as the L1 data on B2 and C1-levels. The *L2 norm* has as high precision as the L1 data had at its best, with 98% precision on A1, B1, B2. However the precision drops on A2 to 96% and also on C1 where it is on the same level as L1 and L2 orig, 97%.

Dependency Annotation Our results show that dependency annotation is less reliable even for L1, with a preserved tendency of quality loss on *L2 orig* as in the lemmatization and POS-annotation. In this case, however, the performance drops by 11 points, from 74.5% to 63%. Normalization improves performance of the Sparv-tool by 6 points, from 63% to 69%. Level of proficiency seems to have a direct effect on the improvement of annotation of *L2 orig*, and for dependency relations also of *L2 norm* except for C1 level. The results for *L1 Coctail* are in line with previous results reporting a LAS score of 78.39 on L1 text (Berdicevskis, 2020a) in automatic evaluation of dependency-relation annotation with Sparv (v.3.0).

We see that our general assumptions are confirmed: the performance of the automatic annotation on learner essays (*L2 orig*) has lower accuracy than on native (*L1 Coctail*) or normalized (cf corrected) (*L2 norm*) texts, even though only marginally for lemmatization and POS tagging. This echoes the results obtained in the automatic evaluation of the Sparv POS-tagging on in-domain L1 texts versus out-of-domain Internet texts (accuracy 0.98 vs 0.93) (Berdicevskis, 2020b) and partially for dependency annotation (Berdicevskis, 2020a). While dependency relation is only moderate in quality, the automatic lemmatization and POS tagging are reliable enough to base further generalizations about L2 development.

4.2 Automatic Detection of MWEs

The purpose of the MWE check in our experiment was to find out whether MWEs: (1) were correctly identified; (2) failed to be identified; (3) were incompletely identified; or (4) were incorrectly identified in the different datasets. Table 3 shows precision, recall and F1 score per resource, as well as a breakdown over the different CEFR levels. These values are calculated relative to the number of automatically and manually identified MWEs (\approx the total correct number of MWEs) and not on a token basis. Numbers are averaged over the two annotators, with standard deviation indicated in parentheses.

F1-scores in Table 3 follow the same tendency as the features described earlier: the pipeline performs best on *L1 Coctail*, the performance drops on *L2 orig* (in this case by 12 points), and improves on *L2 norm* (by 4 points). We cannot see any clear tendency across proficiency levels, the increases and decreases seem to be idiosyncratic and depend on other factors than levels of proficiency, e.g. text genres, topic or task types. Still the results in Table 3 indicate that we can expect that out of 10 MWEs, 7–8 are correctly captured, 2–3 are missed and a small percentage of noise is introduced in the form of suggestions of MWEs that are not actually in the text or that are incomplete MWEs. In nine of the missed cases (45%) an MWE entry is also missing in the Saldo lexicon. However, there are also cases where the MWEs did exist in Saldo but were still missed. All in all, results of this evaluation suggest that we can trust the automatic MWE identification, even though we need to be aware of possible misses.

4.3 Automatic Word Sense Disambiguation (WSD)

The goal of this check was to find out how often: (1) sense was correctly identified; (2) no sense was assigned at all; (3) a lemmagram for the correct sense was missing in Saldo; and (4) the correct sense was missing in Saldo. Table 4 shows the results of the WSD annotation checks. In all three datasets the accuracy of WSD is high, with very slight fluctuations between the datasets. Counter to our expectations, we do not see any radical improvement in performance following normalization of L2 data, nor is there any distinct tendency for poorer WSD quality on the lower proficiency levels. The check shows that some senses are missing in Saldo; sometimes even lemmagrams are missing. Most challenging are function words, like *som*, *mången*, *än* ‘as, much, yet’, that have very few (sense-based) entries in Saldo, and often in combination with a POS that does not match POS tagging based on SUC. For example, for the word *som* ‘as, like’, the SUC taxonomy used in Sparv contains two POS - *conjunction* (KN) and *relative pronoun* (HP), whereas in Saldo, *som* is listed as *subjunction* (SN) and *adverb* (AB), leaving no overlap

Resource	Identified	Correct	Partial	Incorrect	Missed	Precision	Recall	F1
L1 Coctail	59	50.5 (1.5)	4.0 (1.0)	4.5 (0.5)	13.5 (3.5)	0.85 (0.02)	0.79 (0.04)	0.82 (0.03)
A1	8	6.5 (1.5)	1.0 (1.0)	0.5 (0.5)	1.0 (1.0)	0.81 (0.18)	0.85 (0.14)	0.83 (0.16)
A2	7	6.0 (0.0)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.85 (0.0)	0.85 (0.0)	0.85 (0.0)
B1	18	16.5 (0.5)	0.0 (0.0)	1.5 (0.5)	4.5 (0.5)	0.91 (0.02)	0.78 (0.02)	0.84 (0.02)
B2	17	14.0 (1.0)	2.0 (0.0)	1.0 (1.0)	6.5 (1.5)	0.82 (0.05)	0.68 (0.03)	0.74 (0.00)
C1	9	7.5 (0.5)	1.0 (0.0)	0.5 (0.5)	0.5 (0.5)	0.83 (0.05)	0.93 (0.06)	0.88 (0.05)
L2 orig	81	56.0 (1.0)	1.0 (0.0)	24.0 (1.0)	21.0 (2.0)	0.69 (0.01)	0.72 (0.01)	0.70 (0.00)
A1	5	5.0 (0.0)	0.0 (0.0)	0.0 (0.0)	4.5 (1.5)	1.00 (0.0)	0.53 (0.08)	0.69 (0.07)
A2	4	3.0 (0.0)	0.0 (0.0)	1.0 (0.0)	3.5 (0.5)	0.75 (0.0)	0.46 (0.03)	0.57 (0.02)
B1	15	5.0 (0.0)	0.0 (0.0)	10.0 (0.0)	5.5 (0.5)	0.33 (0.0)	0.47 (0.02)	0.39 (0.00)
B2	22	15.0 (0.0)	0.0 (0.0)	7.0 (0.0)	3.5 (0.5)	0.68 (0.0)	0.81 (0.02)	0.74 (0.00)
C1	35	28.0 (1.0)	1.0 (0.0)	6.0 (1.0)	4.0 (1.0)	0.80 (0.02)	0.87 (0.02)	0.83 (0.00)
L2 norm	98	68.5 (1.5)	3.0 (2.0)	26.5 (0.5)	17.5 (0.5)	0.69 (0.01)	0.79 (0.00)	0.74 (0.01)
A1	8	6.5 (0.5)	0.5 (0.5)	1.0 (0.0)	4.5 (0.5)	0.81 (0.06)	0.59 (0.04)	0.68 (0.05)
A2	8	6.0 (0.0)	0.0 (0.0)	2.0 (0.0)	3.5 (0.5)	0.75 (0.0)	0.63 (0.03)	0.68 (0.01)
B1	23	13.5 (0.5)	1.5 (1.5)	8.0 (2.0)	3.5 (1.5)	0.58 (0.02)	0.79 (0.07)	0.67 (0.04)
B2	33	21.0 (1.0)	0.0 (0.0)	12.0 (1.0)	2.0 (0.0)	0.63 (0.03)	0.91 (0.00)	0.74 (0.02)
C1	26	21.5 (0.5)	1.0 (0.0)	3.5 (0.5)	4.0 (1.0)	0.82 (0.01)	0.84 (0.03)	0.83 (0.02)

Table 3: Number of correctly identified MWEs including precision, recall and F1 score: Averages (and standard deviations)

between the two resources. Besides, the sense inventory for such words is too limited in Saldo to cover all the possible contexts where they are used, which we can see in the fact that Svensk ordbok (SO) lists six senses for *som*.

Checking the quality of the automatic WSD on our three datasets has shown that we can expect that in 80–90 percent of the cases the word sense is correctly assigned. Despite the fact that the WSD in Sparv is not bullet-proof, we consider it reliable enough to build our vocabulary resource (L2 lexical profile) on the sense level using lemma+POS+sense as our main entry.

For WSD, a frequently used baseline is the *most frequent sense* baseline which assigns the most frequent sense observed in the training data to each word (Mihalcea, 2007, p. 123). Saldo senses are not ordered by frequency (Borin et al., 2013), thus such a baseline is difficult – albeit not impossible – to calculate (before calculating frequencies, one would need to clarify and justify which corpora to use, etc.). Sense distinctions in Saldo simply indicate that there is a *difference* in sense (Borin et al., 2013). We therefore calculate a simplified version of the *most frequent sense* baseline – the *first sense* baseline – which assigns each word the *first* sense in Saldo. Table 5 shows the number of correct word senses according to the baseline calculation, in comparison with the annotations by annotators 1 and 2, the total number of tokens per dataset, and the mean accuracy and standard deviation. With an average accuracy of about 75%, this baseline is clearly outperformed by the WSD in Sparv. The results for WSD by Sparv are, thus, very encouraging.

5 Qualitative Analysis

In this section we take a closer look at the POS-annotation and the dependency relations in the three datasets. Grammatical annotation such as this can prove very useful both in research and applications in relation to L2 acquisition, but we need to know exactly which tags that are reliable enough.

5.1 Qualitative Analysis of the POS Check

Most of the POS have a precision between 1–0.9 in most of the three datasets and according to both annotators, hence POS-tagging generally provides a very good basis for both research and applications

	# tokens excl punct	Correct sense	Incorrect sense	No sense	Lemgram missing in Saldo	Sense missing in Saldo	Accuracy (correct/ total)
L1 Coctail	1900	1619.5 (66.5)	192.0 (61.0)	46.5 (3.5)	25.0 (1.0)	17.0 (1.0)	0.85 (0.03)
A1	434	399.5 (5.5)	26.5 (4.5)	2.0 (0.0)	5.0 (2.0)	1.0 (1.0)	0.92 (0.01)
A2	101	84.0 (7.0)	15.5 (6.5)	0.5 (0.5)	1.0 (1.0)	0.0 (0.0)	0.83 (0.07)
B1	554	466.0 (22.0)	58.5 (21.5)	14.0 (2.0)	6.5 (1.5)	9.0 (0.0)	0.84 (0.04)
B2	488	409.0 (15.0)	47.5 (13.5)	18.5 (1.5)	8.0 (0.0)	5.0 (0.0)	0.84 (0.03)
C1	324	262.0 (17.0)	44.0 (15.0)	11.5 (0.5)	4.5 (0.5)	2.0 (2.0)	0.81 (0.05)
L2 orig	3635	3000.0 (178.0)	326.5 (163.5)	201.5 (13.5)	25.5 (3.5)	81.5 (2.5)	0.83 (0.05)
A1	301	243.5 (11.5)	33.0 (10.0)	22.5 (2.5)	1.0 (0.0)	1.0 (1.0)	0.81 (0.04)
A2	481	405.5 (14.5)	39.5 (13.5)	27.5 (0.5)	4.0 (1.0)	4.5 (0.5)	0.84 (0.03)
B1	814	657.0 (51.0)	78.5 (42.5)	68.0 (7.0)	5.0 (0.0)	5.5 (1.5)	0.81 (0.06)
B2	886	737.5 (39.5)	76.5 (36.5)	45.0 (1.0)	4.5 (2.5)	22.5 (0.5)	0.83 (0.04)
C1	1153	956.5 (61.5)	99.0 (61.0)	38.5 (2.5)	11.0 (0.0)	48.0 (2.0)	0.83 (0.05)
L2 norm	3565	2963.5 (109.5)	372.5 (108.5)	123.5 (1.5)	30.5 (2.5)	75.0 (3.0)	0.83 (0.03)
A1	323	271.5 (7.5)	40.5 (8.5)	6.0 (0.0)	1.0 (0.0)	4.0 (1.0)	0.84 (0.02)
A2	499	426.0 (5.0)	45.0 (6.0)	15.0 (0.0)	7.5 (0.5)	5.5 (0.5)	0.85 (0.01)
B1	852	718.5 (36.5)	92.0 (33.0)	23.5 (2.5)	6.5 (0.5)	11.5 (0.5)	0.84 (0.04)
B2	1159	966.5 (32.5)	107.0 (31.0)	54.0 (1.0)	7.5 (2.5)	24.0 (0.0)	0.83 (0.03)
C1	732	581.0 (28.0)	88.0 (30.0)	25.0 (0.0)	8.0 (0.0)	30.0 (2.0)	0.79 (0.04)

Table 4: Overview of the automatic sense annotation in the three datasets: Averaged counts (and standard deviation)

Resource	Correct (Annotator 1)	Correct (Annotator 2)	Total	Accuracy (std)
L1 COCTAILL	1469	1401	1900	75.52 (1.79)
L2 orig	2800	2588	3635	76.42 (2.34)
L2 norm	2808	2641	3565	74.11 (2.92)

Table 5: WSD first sense baseline

even when based on learner data. *Participles (PC)* have low precision according to both annotators in *L2 orig*, and this is the only time both annotators are in clear agreement that the pipeline is wrong (precision 0.5 and 0.38). However, only eight tokens have been annotated with PC in this dataset so the figures are hardly reliable. Still we know that participles have been problematic in several ways. They can be lemmatized as verbs, adjectives or as their own POS (participles). Both in Saldo (Borin et al., 2013) and in SAG (Teleman et al., 1999) they are treated as an individual POS. A complicating fact, though, is the ability of Swedish past participles to agree with their noun phrase antecedents, which makes their behavior similar to adjectives, e.g. plural *Stolarna* är **täckta** med snö ‘The chairs are **covered** in snow’ vs singular *Bordet* är **täckt** med snö ‘The table is **covered** in snow’. Note also that many adjectives in Swedish are historically derived from participles, e.g. *nöjd* ‘content, happy’ from the verb *nöja sig* ‘be content with’. All these factors combined make distinguishing participles from verbs and adjectives complicated, especially in learner language.

Other POS with low precision by one annotator can have moderate to excellent precision from the other annotator. This is because there are very few tokens which have been tagged with some POS, e.g. *Interjection* – 2 items in *L2 norm* (precision 1 and 0.5) or *Ordinal number* – 5 items in *L1 Coctail* (1 and 0.4). This particular case clearly shows that Saldo can contribute to disagreement between annotators. The two ordinals annotated here, *första* ‘first’ and *tredje* ‘third’, are adjectival lemmas in Saldo. Other ordinals such as *fjärde* ‘fourth’ appear twice in Saldo, once as an adjectival lemma but also as a form in the morphological paradigm for *fyra* ‘four’. Comparing the datasets we see that *första*, *tredje* received no lemma automatically. Both annotators inserted lemmas according to Saldo, but only one of them adjusted the POS-tag from *RO* to *JJ* in agreement with Saldo. When the token is the ordinal *fjärde* this is lemmatized as *fyra* and neither annotator corrects this in *L2 orig*, but in *L2 norm* it is corrected by one to the lemma *fjärde*, but the POS-tag *RO* is left untouched. Disagreements like these

are not errors and can only be avoided by specifying how to treat these in the guidelines. However, even the researchers who are used to working with Saldo were not aware that this was a difference that existed in Saldo and hence could not take it into account in writing the guidelines. Instead, the check has helped to spotlight an inconsistency in Saldo that should be taken into consideration for future developments, but which may be unavoidable to some extent due to differences in how different the ordinal forms are in relation to the cardinal numbers. Since this also appears to bear a direct affect on the success of lemmatization this is clearly of importance to the performance of the pipeline.

Twenty-three tokens in *L2 orig* have been tagged as *particles*, but the precision differs between 0.96 and 0.43. This appears to be related to the definition of particles. They can be seen as a POS of their own, or as e.g. *adverbs* or *prepositions*; and *particle (adverbial)* is sometimes instead seen as a syntactic function. This is the way SAG views them. It seems one annotator followed SAG more closely and this caused disagreement. IAA could here have been improved by a stricter guideline with regards to how to treat *particles*.

Interestingly, *adjectives* also have quite low precision according to one annotator in both *L2 orig* and *L2 norm*. Most cases (61.6%) have been corrected to *determiner*, a category which this annotator seems to be more familiar with than the other annotator and a category which is not normally included in Swedish grammar, nor is it a POS category in SAG (Teleman et al., 1999). Half of the items are the word *många* ‘many’ which is classed as a pronoun by *Svensk ordbok* and also by Saldo, but which is normally used as a prenominal modifier for quantity and hence could according to some theories be classed as *determiner*. However in SUC 3.0 *många* is annotated as *adjective* (76%) or *pronoun*. *Determiner* is used for similar words like *några* ‘some’ or *alla* ‘all’. This is a clear example where it is hard to decide when the pipeline should be considered correct.

To summarize, IAA is easily severely damaged if there are few items that are being evaluated. Lexical items with clear morphological paradigms with many different forms are easier to classify by POS. But lexical items with morphological paradigms which are hardly used for agreement (e.g. *mången* – *många*) and which in comparison show suppletive forms which can be interpreted as independent lemmas (e.g. *mången*, *många* – *flera*, *flest*) the morphological paradigms cause problems for the annotation.

5.2 Qualitative Analysis of the Check of Dependency Relations

Dependencies can be problematic for linguists because – even though they may be familiar with main categories (e.g. subject, object, finite verb) – checking the dependency annotation entails understanding of what should be seen as correct according to that particular dependency grammar (which in our case includes 65 tags). Since the dependency parser has been trained on Talbanken our annotator was instructed to consult the annotations in Talbanken for comparison when uncertain. In addition, she discussed complicated cases in detail with one of the researchers. Another complicating factor turned out to be that L1 data included some lyrics and poems which were difficult for the parser since sentences were not marked as usual. Similar problems can often be seen in L2 language at low proficiency levels. Unfortunately, few of the dependency labels have a precision above 0.9, only eight in *L1 Coctail*, four in *L2 orig* and four in *L2 norm*. In addition, several labels have been assigned to very few tokens and hence the accuracy is not really reliable as shown above. There are only three categories with a precision between 0.9–0.95 and 39 tokens or more.

It is only *Infinitive Verb phrase minus infinitive marker* (IF) and *Negation adverbial* (NA) that have a precision of 0.9 or more in all the three data sets. In *L1 Coctail* this is based on very little data, but in L2 datasets it is based on 50–63 tokens which is reassuring. Unfortunately, these particular dependency labels do not give that much additional power to L2 research or applications since they are highly correlated with specific words or morphological forms, the negation *inte* ‘not’ and the infinitive. The correct NA-labels are always correlated with the lemma *inte*. Out of all NA in Talbanken $697/742 = 94\%$ are attached to *inte*. And out of all the *inte* in Talbanken $697/720 = 97\%$ are NA. Of course looking at the actual dependency tree it is of interest to see that this dependency relation is related to the correct nodes in the tree since this can affect the semantic interpretation of the sentence.

Nominal adjectival pre-modifiers (AT) are rather well annotated in all datasets. In *L2 orig* the precision is at its lowest at 0.85 (based on 100 tokens) and increases to 0.92 (based on 83) in *L2 norm*. Neither as high as the L1 data, precision 0.95, but the L1 data is based on only 39 tokens and could therefore be seen as less certain. *AT* are interesting for L2 acquisition in relation to both agreement and definiteness. Hence these are important to capture for assessment purposes, CALL and research. Moreover, being able to use extended noun phrases with adjectival premodifiers can be seen as a first step to increased proficiency even if the forms are incorrect.

In comparison to the pre-nominal adjectival modifiers it would be interesting to also be able to catch predicative complements well since they also show agreement to some extent and this type of agreement is more difficult to a learner according to Pienemann's processability theory (Pienemann and Håkansson, 1999) since it crosses phrase boundaries. *Predicative complements* have somewhat lower precision. It is moderate for *subjective predicative complements (SP)* and there is a clear improvement from *L2 orig* to *L2 norm*, but interestingly it does not quite reach L1 precision.

Finally, one last dependency which receives reasonably good scores and is also based on a fair number of tokens is *determiner (DT)*, 0.83 (*L1 Coctail*, 196), 0.87 (*L2 orig*, 357) and 0.92 (*L2 norm*, 339). It is interesting that here L1 has the lowest precision and we see a clear improvement from *L2 orig* to *L2 norm*. *DT* has been attached to tokens which vary quite a lot. Their POS-tags include: conjunctions (KN), determiners (DT), adjectives (JJ), nouns (NN).

6 Conclusions

To summarize, we have seen that lemmatization, POS-tagging and word sense disambiguation are the least sensitive to being applied to non-native data instead of L1. Most affected are dependency annotation and identification of multi-word expressions. All of the annotation steps perform better when applied to normalized learner data instead of the original.

Comparing a non-standard text to a standard text is complicated, and such an evaluation is affected by the type of texts which are used in the evaluation, including the levels of text complexity and the proficiency levels of the essay writers. One complication in evaluating learner texts is that mistakes can be on many different levels. A word might have been used in the wrong context but annotated correctly based on the morphological principles, disregarding semantic and syntactic principles.

Despite the challenges and varying results per linguistic features, we find that our hypotheses have been generally confirmed:

1. Pipelines trained on standard language do not perform equally well on non-standard deviating language. The performance drop varies between different linguistic features, and in certain cases it is relatively negligible (e.g. lemmatization and POS tagging).
2. We have shown that normalization of the learner language improves the performance of the automatic pipeline for all linguistic features, but sometimes only marginally.
3. Proficiency levels have no systematic influence on the performance of the automatic pipeline, apart from in *L2 orig* where there are improvements for each of the linguistic features with growing proficiency levels. This may be due to the fact that automatic pipelines are more sensitive to incorrect language typical of *L2 original* data than to length of the sentences, lexical and syntactic complexity in normlike written texts of different genres and levels.

All in all, the results of our evaluation are very encouraging, especially with regards to lemmatization, POS tagging, and word sense disambiguation. MWE identification seems to be a cognitively more challenging task. Further, we have strong indications that automatic dependency relation annotation is relatively unreliable with the exception of the labels IF, NA, AT, DT, SS and ROOT, and to some extent FS if we disregard the low precision in L1 data because it is based on so few instances. We should therefore be selective in which categories we use for theoretical generalizations and practical implementations. However, the new version of the Sparv pipeline may perform reliably enough for our purposes for all categories.

Acknowledgements

This work has been supported by a research grant from the Swedish Riksbankens Jubileumsfond *Development of lexical and grammatical competences in immigrant Swedish*, P17-0716:1, and by *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions. We also wish to thank the anonymous reviewers for their valuable comments on a previous version.

References

- Alfter D., Lindström Tiedemann T., and Volodina E. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. *Northern European Journal of Language Technology*.
- Berdicevskis A. 2020a. Choosing a new dependency parser for Sparv. Technical report, University of Gothenburg, Department of Swedish, 2020-06-03.
- Berdicevskis A. 2020b. Choosing a new POS-tagger for Sparv: Update. Technical report, University of Gothenburg, Department of Swedish, 2020-05-12.
- Bhalla V. and Klimcikova K. 2019. Evaluation of automatic collocation extraction methods for language learning. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*.
- Borin L., Forsberg M., and Lönngrén L. 2013. SALDO: a touch of yin to WordNet’s yang. *Language Resources and Evaluation*, 47(4):1191–1211.
- Borin L., Forsberg M., Hammarstedt M., Rosén D., Schäfer R., and Schumacher A. 2016. Sparv: Språkbanken’s corpus annotation pipeline infrastructure. In *Proceedings of Swedish Language Technology Conference (SLTC)*. Umeå University.
- Borin L. 2021. Multiword expressions—a tough typological nut for swedish framenet++1. In Dannells D., Borin L., and Friberg Heppin K., editors, *The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications*, volume 14, pages 221–262. John Benjamins Publishing Company.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.
- De Marneffe M.-C., Dozat T., Silveira N., Haverinen K., Ginter F., Nivre J., and Manning C. D. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.
- Ejerhed E., Källgren G., and Brodda B. 1997. Stockholm Umeå Corpus version 1.0, SUC 1.0. *Department of Linguistics, Umeå University*.
- Gardner D. 2007. Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied linguistics*, 28(2):241–265.
- Gustafson-Capková S. and Hartmann B. 2006. Manual of the stockholm umeå corpus version 2.0. *Unpublished Work*.
- Halácsy P., Kornai A., and Oravecz C. 2007. HunPos—an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, United States. Association for Computational Linguistics.
- Krippendorff K. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.
- Krivanek J. and Meurers D. 2013. Comparing rule-based and data-driven dependency parsing of learner language. *Computational dependency theory*, 258:207.
- Kübler S., McDonald R., and Nivre J. 2009. Dependency parsing. *Synthesis lectures on human language technologies*, 1(1):1–127.
- Ljunglöf P., Zechner N., Nieto Piña L., Adesam Y., and Borin L. 2019. Assessing the quality of Språkbanken’s annotations. Technical report, University of Gothenburg, Department of Swedish.
- Meurers D. and Wunsch H. 2010. Linguistically annotated learner corpora: Aspects of a layered linguistic encoding and standardized representation. *Proceedings of Linguistic Evidence*, pages 1–4.
- Mihalcea R. 2007. Knowledge-based methods for WSD. In *Word sense disambiguation*, pages 107–131. Springer.
- Nieto Piña L. 2019. *Splitting rocks: Learning word sense representations from corpora and lexica*. PhD Thesis, Data Linguistica 30.
- Nilsson J., Hall J., and Nivre J. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. *Proceedings from the special session on treebanks at NoDaLiDa 2005*, pages 119–132.
- Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kübler S., Marinov S., and Marsi E. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95.
- Nivre J., Megyesi B., Gustafson-Capková S., Salomonsson F., and Dahlqvist B. 2008. Cultivating a Swedish treebank. *Resourceful language technology: Festschrift in honor of Anna Săgvall Hein*, pages 111–120.

- Ott N. and Ziai R. 2010. Evaluating dependency parsing performance on german learner language. In *Proceedings of the ninth international workshop on treebanks and linguistic theories*, volume 9, pages 175–186. NEALT Tartu.
- Paquot M. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1):121–145.
- Petrov S., Das D., and McDonald R. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.
- Pienemann M. and Håkansson G. 1999. A unified approach toward the development of swedish as l2: A process-ability account. *Studies in second language acquisition*, 21(3):383–420.
- Qi P., Zhang Y., Zhang Y., Bolton J., and Manning C. D. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- Rubin R. 2021. Assessing the impact of automatic dependency annotation on the measurement of phraseological complexity in l2 dutch. *International Journal of Learner Corpus Research*, 7(1):131–162.
- Rudebeck L., Sundberg G., and Wirén M. 2021. SweLL normalization guidelines. Technical report, GU-ISS Research report series, Department of Swedish, University of Gothenburg. <http://hdl.handle.net/2077/69432>.
- Sag I. A., Baldwin T., Bond F., Copestake A., and Flickinger D. 2002. Multiword expressions: A pain in the neck for nlp. In *Conference on intelligent text processing and computational linguistics*. Springer.
- Štindlová B., Rosen A., Hana J., and Škodová S. 2012. CzeSL—an error tagged corpus of Czech as a second language. In *Corpus data across languages and disciplines*. Peter Lang.
- Teleman U., Hellberg S., Andersson E., et al. 1999. Svenska akademiens grammatik. *Arkiv*, page 233.
- Volodina E., Pilán I., Eide S. R., and Heidarsson H. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the 3rd Workshop on NLP for Computer-Assisted Language Learning*. Linköping University Press.
- Volodina E., Pilán I., Enström I., Llozhi L., Lundkvist P., Sundberg G., and Sandell M. 2016. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia.
- Wirén M., Matsson A., Rosén D., and Volodina E. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *Proceedings of CLARIN 2018*.
- Wray A. 2005. *Formulaic language and the lexicon*. Cambridge University Press.

A Appendix: POS taxonomy

Kod ‘Code’ <i>Ordklass</i> ‘Word class’	Svensk term ‘Swedish term’	Engelsk term ‘English term’
AB	Adverb	Adverb
DT	Determinerare, bestämningsord	Determiner
HA	Frågande/relativt adverb	Interrogative/Relative Adverb
HD	Frågande/relativt bestämning	Interrogative/Relative Determiner
HP	Frågande/relativt pronomen	Interrogative/Relative Pronoun
HS	Frågande/relativt possessivuttryck	Interrogative/Relative Possessive
IE	Infinitivmärke	Infinitive Marker
IN	Interjektion	Interjection
JJ	Adjektiv	Adjective
KN	Konjunktion	Conjunction
NN	Substantiv	Noun
PC	Particip	Participle
PL	Partikel	Particle
PM	Egennamn	Proper Noun
PN	Pronomen	Pronoun
PP	Preposition	Preposition
PS	Possessivuttryck	Possessive
RG	Räkneord: grundtal	Cardinal Number
RO	Räkneord: ordningstal	Ordinal Number
SN	Subjunktion	Subjunction
UO	Utländskt ord	Foreign Word
VB	Verb	Verb

Table 6: SUC-based part of speech categories (POS code, Swedish term, English term).

B Appendix: DepRel taxonomy

MAMBA Categories			
Tag	Meaning	Tag	Meaning
++	Coordinating conjunction	JR	Second parenthesis
+A	Conjunctive adverbial	JT	Second dash
+F	Coordination at main clause level	KA	Comparative adverbial
AA	Other adverbial	MA	Attitude adverbial
AG	Agent	MS	Macrosyntagm
AN	Apposition	NA	Negation adverbial
AT	Nominal (adjectival) pre-modifier	OA	Object adverbial
CA	Contrastive adverbial	OO	Direct object
DB	Doubled function	OP	Object predicative
DT	Determiner	PL	Verb particle
EF	Relative clause in cleft	PR	Preposition
EO	Logical object	PT	Predicative attribute
ES	Logical subject	RA	Place adverbial
ET	Other nominal post-modifier	SP	Subjective predicative complement
FO	Dummy object	SS	Other subject
FP	Free subjective predicative complement	TA	Time adverbial
FS	Dummy subject	TT	Address phrase
FV	Finite predicate verb	UK	Subordinating conjunction
I?	Question mark	VA	Notifying adverbial
IC	Quotation mark	VO	Infinitive object complement
IG	Other punctuation mark	VS	Infinitive subject complement
IK	Comma	XA	Expressions like "så att säga" (so to speak)
IM	Infinitive marker	XF	Fundament phrase
IO	Indirect object	XT	Expressions like "så kallad" (so called)
IP	Period	XX	Unclassifiable grammatical function
IQ	Colon	YY	Interjection phrase
IR	Parenthesis	New Categories	
IS	Semicolon	CJ	Conjunct (in coordinate structure)
IT	Dash	HD	Head
IU	Exclamation mark	IF	Infinitive verb phrase minus infinitive marker
IV	Nonfinite verb	PA	Complement of preposition
JC	Second quotation mark	UA	Subordinate clause minus subordinating conjunction
JG	Second (other) punctuation mark	VG	Verb group

Table 7: MAMBA categories for annotation of dependency relations (DepRel code, English term).

C Appendix: Statistics of the three datasets

Dataset	Level	# sent.	# tokens excl.punct
L1 Coctail	15 texts	196	1900
	A1	57	434
	A2	19	101
	B1	57	553
	B2	32	488
	C1	31	324
L2 orig	15 texts	287	3635
	A1	42	301
	A2	60	481
	B1	61	814
	B2	63	886
	C1	61	1153
L2 norm	15 texts	306	3565
	A1	52	323
	A2	60	499
	B1	65	852
	B2	64	1159
	C1	65	732
Total	45 texts	789	9100

Table 8: Statistics over the three datasets

Flexible Metadata Schemes for Research Data Repositories The Common Framework in Dataverse and the CMDI Use Case

Jerry de Vries
DANS-KNAW
The Netherlands

`jerry.de.vries@dans.knaw.nl`

Vyacheslav Tykhonov
DANS-KNAW
The Netherlands

`vyachesav.tykhonov@dans.knaw.nl`

Andrea Scharnhorst
DANS-KNAW
The Netherlands

`andrea.scharnhorst@dans.knaw.nl`

Eko Indarto
DANS-KNAW
The Netherlands

`eko.indarto@dans.knaw.nl`

Mike Priddy
DANS-KNAW
The Netherlands

`mike.priddy@dans.knaw.nl`

Femmy Admiraal
DANS-KNAW
The Netherlands

`femmy.admiraal@dans.knaw.nl`

Abstract

In this paper we present an approach called Common Framework, which addresses issues of interoperability and flexibility of metadata schemes as developed by specific scientific communities, and as later supported by domain and cross-domain data repositories. The approach was triggered by a very concrete use case, namely the question how to expose Component Metadata Infrastructure (CMDI) metadata, stored in computational linguistics datasets in the DANS-EASY archive, for discovery services. The work in CLARIN to push further for the development of CMDI into a standard (ISO 24622-1:2015, ISO 24622-2:2019) forms part of the background of the use case. We used the Dataverse platform to deliver proof of concepts for various elements of the Common Framework, including the recommendation of standardised elements for Dataverse instances in CLARIN. At the core of the Common Framework is a design which envisions an interaction between different microservices, possibly also hosted by various service providers. Mechanisms of semantic mapping are used throughout a pipeline which starts at a set of existing metadata standards and values at a digital research data repository (Extraction) and their analysis. This leads to an alignment of these metadata standards with others standards (Transformation) and proposes enrichments to be used by other service providers but also to be imported back to the original source (Load). Some modules applied along this pipeline are discussed in detail, together with the challenges this specific use case entails. At the same time, we also stress generic aspects, as we are convinced that this approach can also be applied in other settings, other archival platforms and other domain specific metadata schemes. The high-level goal of this exploration is to explore ways to make research data collections FAIR (Findable, Accessible, Interoperable and Re-usable), and in particular interoperable and re-usable, while preserving the rigour of domain specific indexing practices.

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

Jerry de Vries, Vyacheslav Tykhonov, Andrea Scharnhorst, Eko Indarto, Mike Priddy and Femmy Admiraal 2022. Flexible Metadata Schemes for Research Data Repositories. The Common Framework in Dataverse and the CMDI Use Case. *Selected papers from the CLARIN Annual Conference 2021*. Ed. by Monica Monachini and Maria Eskevich. Linköping Electronic Conference Proceedings 189, pp. 168–180. DOI: <https://doi.org/10.3384/9789179294441>

1 Introduction

Research data repositories are increasingly expected to operate together. Standardisation and alignment of metadata schemes used to describe (index) datasets are a precondition for any platform to work (for example see <https://datacite.org>). At the same time, data repositories usually serve specific knowledge domains, and have tailored their indexing practices towards those communities. In short, there is a tension between serving one or few communities in a very rigorous manner and being integratable into cross-domain platforms (see Figure 1).

The tension between specificity of metadata schemes and a genericity which enables interoperability is nothing new (e.g., Guéret et al., 2013). We see similar debates around the emergence of universal classifications in the bibliographic domain at the beginning of the twentieth century (e.g., Dewey and UDC) (McIlwaine, 2010); reinforced with the introduction of automatization in classification and indexing (Svenonius, 2000); and reappearing in a different shape with the emergence of web services. Currently, (traditional) phrases as *crosswalks*, *alignment*, *catalogues* mark the quest for interoperability in the growing universe of domain specific ontologies, classifications, thesauri which become semantic artefacts when living in the web (Hugo et al., 2020; European Commission, Directorate-General for Research and Innovation et al., 2021).

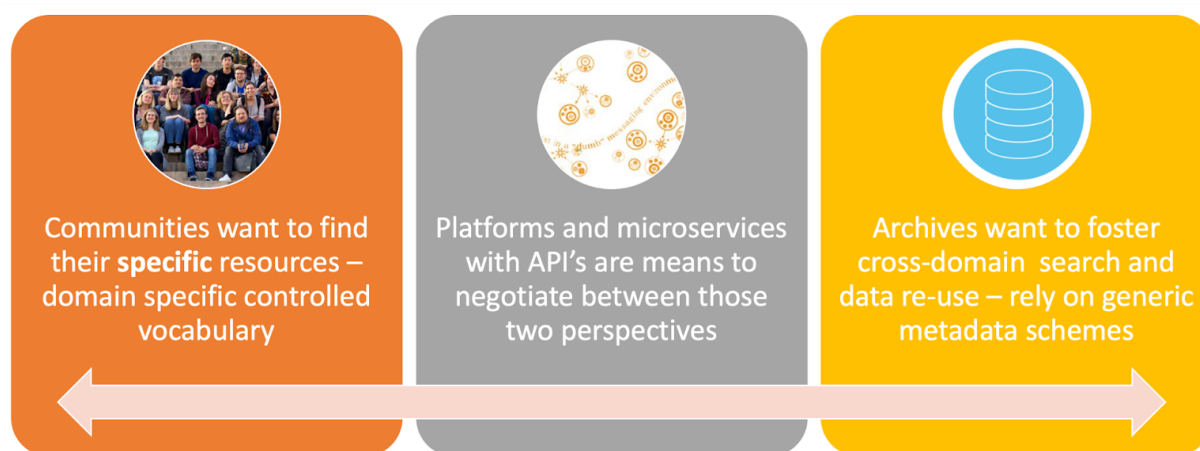


Figure 1. Tension between specific resources and generic metadata schemes

While almost trivial, one cannot overemphasise the fact that the organisation of knowledge, and of systems and practices around it, is deeply contextualised, and depends on the concrete purpose for such knowledge orders to emerge (Smiraglia and Scharnhorst, 2021). So, it cannot be a surprise that each knowledge domain, each scientific field or speciality according to its epistemic frameworks and research perspectives develops its own specific organisation of knowledge. One could even say that research knowledge organisation systems have an element of being intrinsically not-interoperable. This is the curse to the virtue of deepening our knowledge in an ever more differentiated and specialised knowledge universe (Scharnhorst and Smiraglia 2021). Additionally, even in one domain the organisation of knowledge once achieved does not remain static (e.g., Tennis, 2018).

Formulated in terms of stability and volatility one could say that research itself naturally leads to changed knowledge. Volatility in this sense is what research is about. The organisation of new knowledge content needs to be a bit more stable to enable communication and knowledge transfer across all involved actors in a domain. But, with ever newly emerging knowledge also this domain organisation will need to change. If it comes to knowledge exchange across domains the reach/scale of interaction is bigger, and so stability is even more important, just to allow all parts of the information system to align with each other. One could also say the larger the (information) systems are, in which knowledge is produced and exchanged, the slower adaptation needs to take place to prevent a disconnect of parts of the system. However, at the end, even standardisation is relative, time-dependent and operates on different time scales.

It is these different timescales of change we refer to if we talk about flexible metadata schemes for repositories. In this paper, we look more closely into the negotiation between scientific communities

and repositories, using one specific case: the DANS Long-term preservation archive EASY¹ and the CMDI metadata framework of the computational linguistics community (Goosen et al., 2015). But we developed our approach in a way that is also applicable for other metadata schemes: hence the choice of the name *Common Framework*.

A central aspect in our design is the exploration of moving from one application (in our case Fedora as the repository software behind EASY) to a modular approach of a coherent set of microservices working together, making the system more flexible towards the future. On a general level this is in line with thinking of infrastructures as consisting of networked ‘microservices’ (Wang, Y et al., 2021). We chose Dataverse as our main application to execute several workflows due to its active open development community², our own in-depth experiences with developing Dataverse microservices in various projects, the DANS experience as host of a Dataverse platform service for Dutch Higher Education repositories, and the fact that the CLARIN community has an instance of Dataverse in Norway with the colleagues of which we have already collaborated (Conzett et al., 2020). Moreover, Dataverse has already responded to the need of flexible metadata schemes by offering both a standard, common core set of metadata called Citation Block³ and the possibility to extend this core set with custom fields defined as a discipline specific metadata block⁴.

Our use case unfolds around a concrete pipeline - called the ETL pipeline (Extract-Transform-Load). We start with a CLARIN and Oral History collection, indexed on the dataset level with a specific metadata standard, and with more specific indexing information which can be found in a specific CMDI metadata file as part of the datasets in this collection. We call this phase *Extract*. We extract and analyse both schemas and values, with the aim to prepare alignments. These alignments are executed in the next phase (*Transform*). At the end, we discuss how enriched information can be feedback to the source repository as well as made available for other service providers (*Load*).

While we departed from DANS-EASY and the ‘CMDI use case’⁵, during our exploration it became obvious that various modules we developed as proof of concepts can also be applied to other settings: other metadata schemes, other problems of alignments and so one. This gradually led to the emergence of the *Common Framework*. The recommendation of standardised elements for Dataverse instances in CLARIN then becomes a special example of this *Common Framework*. At the core of the *Common Framework* is a design which envisions an interaction between different microservices, possibly also hosted by various service providers. Mechanisms of semantic mapping are the cornerstones of the framework.

In the next section, we unfold the steps which led to the *Common Framework*, the implementations and challenges we had to respond to.

2 Building a Common Framework

Figure 2 shows the major components upon which the *Common Framework* is built, along the pipeline we introduced above. In the course of the work, we identified two major linking tasks:

- Finding an appropriate ontology for the specific metadata fields (**red block**),
- The prediction and linkage of the appropriate concepts for their values from the list of available controlled vocabularies (**green block**).

The ultimate goal of the whole workflow lies within metadata enrichment, with adding Uniform Resource Identifiers (URIs) for both concepts and their values. For our case this means that we work on increasing the FAIR score of the datasets with associated CMDI metadata.

¹ <https://easy.dans.knaw.nl/>

² <https://dataverse.org/>

³ <https://guides.dataverse.org/en/latest/user/appendix.html>

⁴ <https://guides.dataverse.org/en/4.20/admin/metadatacustomization.html>

⁵ <https://github.com/CLARIAH/CLARIAH-plus/blob/main/use-cases/cases/DANS-cmdi.md>

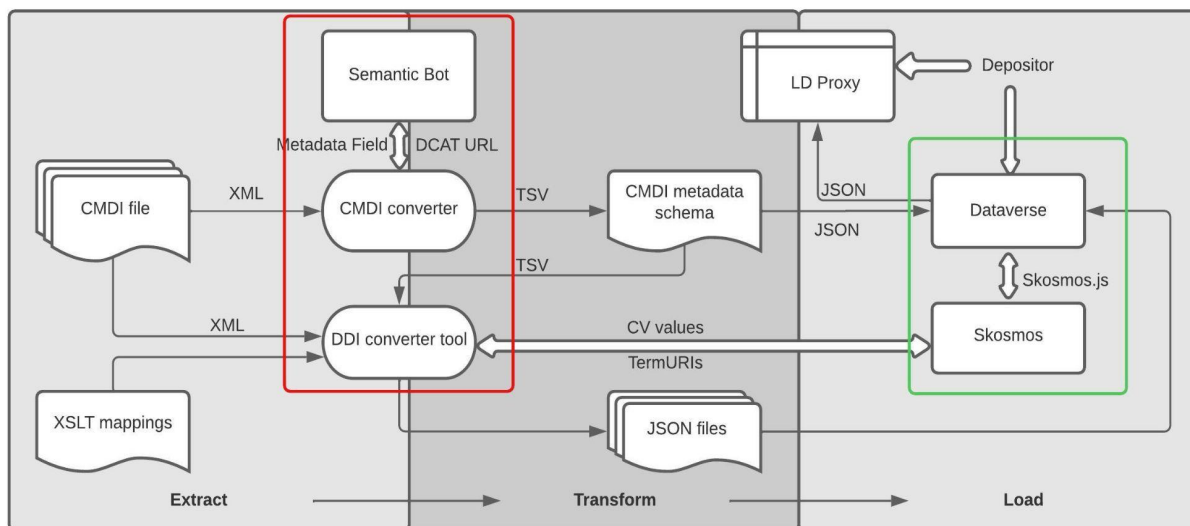


Figure 2. Schematic overview of the workflow (ETL pipeline)

The entire workflow consists of multiple microservices, indicated by blocks in the schematic description, which are separate components that we are building, reusing, or extending. Furthermore, this workflow can be extended with new open-source components (microservices) that already exist and are maintained by other institutions. This setup gives us the flexible opportunity to connect third-party services provided elsewhere. These components are represented as separate blocks in the workflow shown in Figure 2.

Component	Component provider	URL
CMDI converter	DANS-KNAW	https://github.com/DANS-labs/CLARIAH_CMDI
CMDI files	DANS-EASY archive	www.easy.dans.knaw.nl
Dataverse	DANS-KNAW	Internal DANS / Humanities Cluster ⁶ (HuC) development environment
DDI converter tool	DANS-KNAW	https://github.com/IQSS/dataverse-ddi-converter-tool
LD proxy	DANS-KNAW/ KNAW HuC	Resolver: https://github.com/DANS-labs/ld-serverless-resolver LD Proxy: https://github.com/KNAW-HUC/LDProxy
Semantic Bot	Open Data Soft	https://github.com/opendatasoft/semantic-bot
Skosmos	National Library of Finland	https://skosmos.org/
XSLT mappings	KNAW HuC	https://github.com/menzowindhouwer/dataverse2cmdi/blob/main/profile/prof2tsv.xsl

Table 1. Overview of components presented in the ETL pipeline

Examples from the workflow above and presented in Table 1 are: Skosmos as a service (for our proof of concept we have used the Skosmos instance of the National Library of Finland), a shared internal Dataverse instance, a local Semantic Bot installation and a DDI Converter tool. The arrows in the workflow (Figure 2) represent the interaction between the services. By reusing what is already available and eventually connecting it to our own DANS service infrastructure we hope to drive innovation. The integration of microservice-type components in existing workflows contributes, in principle, to their sustainability (sustainability by re-use and integration). In turn, the addition of microservices to existing

⁶ <https://huc.knaw.nl/>

workflows contributes to the innovation of those workflows without the need to rebuild the entire workflow. Having said this, interoperability and maintenance are of course the underlying principles for this design to succeed. Ideally, the objective is to benefit from all the developments happening somewhere and matching these with our current demands and requirements. In the following we demonstrate this for the CMDI use case.

2.1 Describing CMDI use case

As part of the CLARIAH+⁷ project DANS is working on the CMDI use case. The main goal of this use case is to identify all linguistic and oral history datasets containing a CMDI metadata file archived in the DANS-EASY repository and make these datasets visible and harvestable by CLARIN harvesters. What we call a dataset refers to one archival information package (AIP)⁸ in the DANS-EASY archive which comes with a metadata file and can contain several folders or folder structures and/or individual files. The current DANS-EASY archive uses a subset of the Dublin Core Standards and DCMI Terms⁹ as a metadata set to describe a dataset. The difficulty here is that the CMDI based metadata description of the dataset is attached to the dataset as a separate file, in either text or XML format, and thus not included in the EASY search index and also not harvested by CLARIN. The CMDI metadata fields are currently not visible for search interfaces, harvesters and metadata aggregators. The separate CMDI file forms part of the dissemination information package (DIP)¹⁰ for harvesting or download by a researcher, thus it is possible for a user to view the CMDI metadata and potentially use it, however it is not manifest and prominent.

To make the CMDI metadata usable by metadata aggregators and for the discovery of datasets the first task was the identification of all AIPs which have been deposited with CMDI metadata in DANS-EASY. There are different ways to find them however, with the web interface, a general search through all metadata fields revealed (2020) 1096 datasets which use CMDI metadata, and it should be noted that its use is in either the Description metadata field or in the Form metadata field of the Dublin Core Standard. However, as noted above, while the use of CMDI is identified, the web interface cannot be used to automatically search in those CMDI notations as they are in a separate file and thus not indexed for search.

Once identified, the second task was to extract all CMDI specific metadata fields from the CMDI files and perform an analysis on the distribution of fields used and filled. Here, one needs to understand that a) CMDI is a standard, but does not have an obligatory set of core fields and b) the values of CMDI fields could be defined in a schema as free-text, closed vocabularies, specific data types or even regular expressions. We also see in the analysis of the CMDI metadata how this standard performs in the wild, and as expected we see quite some variation (Smiraglia et al., 2013; Odijk, 2016). An alternative to the tool we used could have been the SMC-Browser¹¹ (Durco et al., 2014) of which we were not aware at this time

The analysis was twofold. We first performed a frequency analysis of the CMDI metadata fields in our sample of 1097 records. As this sample was insufficient to draw conclusions on what might be a core set of metadata, we cooperated with CLARIN to do a second analysis executed on the Virtual Language Observatory¹² (VLO) that describes over a million datasets. CLARIN used our tool and performed a frequency and hierarchy analysis to see the distribution of metadata fields used in the CMDI descriptions of the VLO. We must point out here that the frequency analysis did not take into account the CMDI feature that allows different schemas to use different namespaces for the same element. Table 2 shows the result of the frequency analysis. The outcome of both analyses is the basis for further analyses and to identify core elements of CMDI.

⁷ <https://www.clariah.nl/>

⁸ As described in the Reference Model for an Open Archival Information System (OAIS): <https://public.ccsds.org/Pubs/650x0m2.pdf>

⁹ EASY metadata schema: <https://easy.dans.knaw.nl/schemas/md/emd/emd.xsd>

¹⁰ As described in the Reference Model for an Open Archival Information System (OAIS): <https://public.ccsds.org/Pubs/650x0m2.pdf>

¹¹ <https://clarin.oeaw.ac.at/smc-browser/index.html>

¹² <https://vlo.clarin.eu/> - Analysis executed 8th of April, 2020

cmdp:Description; 36774144	cmdp:Code; 18114816
cmdp:AnnotationType; 22291456	cmdp:MimeType; 16102720
cmdp:SizeUnit; 21238016	cmdp:LanguageName; 10574912
cmdp:Number; 21238016	cmdp:TotalSize; 10572928
cmdp:iso-639-3-code; 20817152	cmdp:Name; 10415552

Table 2. Most frequent used metadata fields based on VLO CMDI metadata

The variety in the use of a metadata standard, visible in our use case, has been long debated in the CLARIN community (Windhouwer et al., 2012). CMDI is a metadata framework primarily used for describing digital language resources. There is no obligatory subset of fields required for each CLARIN resource. Of course, this also poses a problem for the CLARIN-wide implementation of a federated search on the metadata in the Virtual Language Observatory, and the community itself is working on this intensively. Moreover, as for any indexing practice, metadata is not always complete or harmonised as the CLARIN community maintains limited mappings for VLO facets, and it is quite time consuming.

Despite these problems, the evaluation of the CMDI metadata from DANS-EASY and the Virtual Language Observatory led to a proposal for a core set of elements¹³. It remains a difficult process to eventually implement such a core set, to which all CLARIN data providers would need to agree to, and the proposal is still under discussion. However, such a draft proposed core set¹⁴ was enough to start a Proof of Concept for the CMDI transformation.

We used the Dataverse platform to implement what became a CMDI metadata block. Firstly, a CMDI converter tool¹⁵ was created and used to extract the CMDI fields from the XML files. Then to create a proposed CMDI metadata block in Dataverse these extracted fields were transformed to a CMDI metadata schema by using Tabs-Separated Values (TSV)¹⁶ file. The Dataverse DDI converter tool¹⁷ was used to convert the TSV files into a JSON file, which could be imported to Dataverse by using the Dataverse API to create the specific CMDI metadata block in Dataverse. With the CMDI metadata block in a Dataverse instance the following step is to load all the extracted CMDI metadata values to the corresponding metadata fields. As the original metadata files are present, a simple field value mapping using the JSON format did the job.

However, as we detail later, there is also a challenge of enriching this metadata by updating each field through linking the value to a corresponding term from a recommended controlled vocabulary. This last step is also part of the *Common Framework*, and thus could be applied to any metadata extraction and transformation. In our CMDI case, this last step (Load) helps us to make CMDI metadata more visible and findable.

2.2 From use case to general framework

The basis for the *Common Framework* has evolved from the desire to achieve a universal Federated Search across multiple data repositories. The current lack of crosswalks and mappings across different metadata schemes and the lack of enriched indexes with values from controlled vocabularies presents a challenge.

The exploration of agile solutions and proof of concepts, in principle of value for different communities, helped us in defining and understanding the problem domain and gave us a clear future perspective: the creation of FAIR metadata and related semantic services. Starting with a conceptual approach for semantic interoperability on the infrastructure level was the basis for finding a common, generic solution suitable for any metadata related use case. We had to make some critical changes in the Dataverse repository core software to implement this conceptual shift, which departs from the traditional

¹³ Working document by Goosen, T., Broeder, D., Windhouwer, M., Köning, A., Labropoulou, P., Conzett, P., Van Uytvanck, D., Oleksy, M., Ohren, O.P., Tykhonov, V., De Vries, J., https://docs.google.com/document/d/1sTgp_rdwE40tMqKqhuUQ2m-FJ74NURNQ1c2IAsC295E/edit

¹⁴ Working document of mapping by Goosen, T., Broeder, D., Windhouwer, M., Köning, A., Labropoulou, P., Conzett, P., Van Uytvanck, D., Oleksy, M., Ohren, O.P., Tykhonov, V., De Vries, J., <https://docs.google.com/spreadsheets/d/1zKR5ErqL3wRX4tOL37110-34jXVP0gNzgU2vFsLrbcl/edit>

¹⁵ https://github.com/DANS-labs/CLARIAH_CMDI

¹⁶ <https://guides.dataverse.org/en/latest/admin/metadatacustomization.html>

¹⁷ <https://github.com/IQSS/dataverse-ddi-converter-tool>

understanding of (meta)data management and leads to semantically driven services. In the following we describe the steps of these implementations and their conceptual importance.

The first conceptual step consisted of the leveraging of the Semantic Metadata API¹⁸ being built for the Dataverse platform by the Global Dataverse Community Consortium¹⁹ (GDCC) as a part of Dataverse's core. DANS has played a critical role in the testing and improvement of this new functionality that was introduced in version 5.6 of Dataverse.²⁰ The format, which follows the OAI-ORE²¹ export recommendations, allows for a standardised transfer of metadata from, and to, external systems without knowledge of the Dataverse specifics, such as metadata block and field storage architecture. More importantly, the Semantic Metadata API allows for the update of metadata fields published in Dataverse, both on the level of the dataset, as well as on the level of the data in the dataset and therefore, could be widely used for metadata enhancement.

The second conceptual step was the extension of the Dataverse API to fully support external controlled vocabularies (Tykhonov, 2021). This functionality was originally developed by DANS for work on a Skosmos framework²² in the Social Science and Humanities Open Cloud²³ (SSHOC) project (Tykhonov et al., 2021) and extended by the GDCC to allow a more generic integration of Wikidata²⁴, ORCID²⁵, MeSH²⁶ and other controlled vocabularies (Tykhonov et al., 2021). A Skosmos implementation helps to get appropriate Simple Knowledge Organisation System (SKOS)²⁷ representations of the relevant controlled vocabularies and serves as a lookup service for metadata values and terms, returning the concept URI²⁸ that could be added to the metadata to enrich the dataset metadata. At the same time this process of metadata standardisation is extremely important for interoperability as the content of the selected concept is being cached in JSON²⁹ and indexed by Dataverse, and therefore, is available in the search interface. The utilisation of concept URIs facilitates users to find the dataset whilst querying in languages other than the original of the deposited dataset. For example, datasets with metadata described in Chinese, Russian or Arabic could be found with English search queries as soon as some of their terms are linked to external multilingual controlled vocabularies. It is important to understand that Dataverse is an open-source data repository and has a global community consortium, consequently all of its community members can potentially obtain and utilise this new semantic-based functionality after they upgrade their running data repository instance to version 5.7 or higher. When we worked on the consensus proposal, we also involved all community members in the process and accepted various comments, contributions, and feedback to make the collaborative solution as generic as possible. It was implemented in a fashion that it could be reused by any data repository system dealing with semantic artefacts (Hugo et al., 2020).

As a result of this work, users can export metadata from Dataverse repositories in the JSON, JSON-LD and other common formats and thus provide a more consistent way for the utilisation of this metadata by developers (and others) to create data-centric applications. Rich metadata descriptions will contain concept URIs with their cached records consisting of JSON export, this information could be indexed by various aggregators and used as a basis for the building of semantic search facilities, basically providing universal federated search across multiple data repositories.

2.3 Flexible Semantic Mapping Framework (SEMAF)

As indicated above, CMDI is a standard for which data providers usually define their own profiles, specifically tailored to their collections, consisting of various components of the CMDI framework. As a consequence, those self-defined metadata schemas in CMDI create complexity (Durco et al., 2018).

¹⁸ <https://guides.dataverse.org/en/latest/developers/dataset-semantic-metadata-api.html>

¹⁹ <https://dataversecommunity.global/>

²⁰ <https://json-ld.org/>

²¹ <https://www.openarchives.org/ore/>

²² <https://github.com/SSHOC/Skosmos>

²³ <https://sshopencloud.eu/>

²⁴ <https://www.wikidata.org/>

²⁵ <https://orcid.org/>

²⁶ <https://www.ncbi.nlm.nih.gov/mesh/>

²⁷ <https://www.w3.org/2004/02/skos/>

²⁸ https://www.wikidata.org/wiki/Wikidata:Data_access

²⁹ <https://www.json.org/json-en.html>

In turn, this has led to increased efforts on the maintenance of mappings for all CMDI fields³⁰. Most of the CMDI mappings are done by the CLARIN community using XSLT transformations (Haaf et al., 2014). To improve the situation, Broeder et al. (2021) proposed the flexible Semantic Mapping Framework (SEMAF) to create, document and publish semantic mappings and cross-walks, linking different semantic artefacts within a particular scientific community and across scientific domains (Myers et al., 2021).

The aim is to keep all mappings in semantic form without taking into account the initial structure and hierarchy of the metadata records, and to reuse those mappings for CMDI, XML, CSV or any kind of input format.

We have developed a proof of concept using the Data Catalog Vocabulary³¹ (DCAT) mappings for CMDI metadata fields to produce a standardised metadata schema where every CMDI metadata field is mapped to a DCAT URL. As soon as this link is established, another process is used to extract all the values from appropriate fields and to update metadata records with the corresponding URIs, helped by the Skosmos look-up service (see Figure 2, phase Load). SKOS is applied here to model thesauri-like resources with simple `skos:broader`, `skos:narrower` and `skos:related` properties.

This approach allows us to load all elements, such as properties and attributes, from CMDI records and build a knowledge graph from them. This solution is suitable for any metadata enhancement task where dataset metadata is being enriched with concept URIs linked from various controlled vocabularies such as Skosmos, Wikidata and others, linking to the appropriate nodes in the knowledge graph. In turn, this knowledge graph could be serialised in formats suitable for the integration with different systems. For example, JSON-LD serialisation works well for Dataverse, TURTLE and RDF/XML serialisations for Apache Jena Fuseki triple stores³² (Tykhonov et al., 2021).

2.4 Using Machine Learning for metadata enrichment

While testing the workflow (schematically depicted in Figure 2) we discovered that the quality of the linking approach, based on the Deterministic (Exact) Matching Method (Shlomo, 2019), and currently available in Skosmos and Wikidata services, was rather poor. In some cases, this lookup process returns a lot of irrelevant candidates as it does not take into account the ambiguity based on the applied context, with the associated possibility of selecting an inappropriate concept URI and creating a false data linkage and an incorrect assertion.

We began to experiment with Machine Learning (ML) in order to add context to improve this workflow and the first results are very promising. We ran a Doccano annotation tool³³, which facilitated collaborative labelling of concepts, over the text of CMDI records and received results with recognised concepts and entities, delivered through our SpaCy³⁴ based Machine Learning pipeline (Figure 3). All annotations are shared between all users as a part of a collaborative effort as, in principle, the labelling of concepts could be improved. Users can also create new labels, highlight them in the text and enrich annotation with comments. After the annotation is complete the ML model is retrained.

Such a collaborative approach enables an increase in the quality of the concepts detected and provides more accurate information about types or classes of concepts, for example, for persons (PERSON), organisations (ORG), dates (DATE), Geo-Political Entity (GPE). It may guide the linking process to create appropriate links between concepts and controlled vocabularies, however, human oversight is required to review ambiguities and changes, for example, in names of persons and places. Future work will incorporate experimenting with the association of concepts, for example, with PERSON to lookup in the ORCID registry (if living), GPE in the Geonames service³⁵, ORG in the Global Research Identifier Database³⁶ (GRID), etc. One should emphasise that this experiment did not check for the ambiguity of, for instance, multiple individuals or places having the same name.

³⁰ <https://www.clarin.eu/content/component-metadata>

³¹ <https://www.w3.org/TR/vocab-dcat-2/>

³² <https://jena.apache.org/documentation/fuseki2/>

³³ <https://github.com/doccano/doccano>

³⁴ <https://github.com/DANS-labs/spacy-DANS>

³⁵ <http://www.geonames.org/>

³⁶ <https://www.grid.ac/>

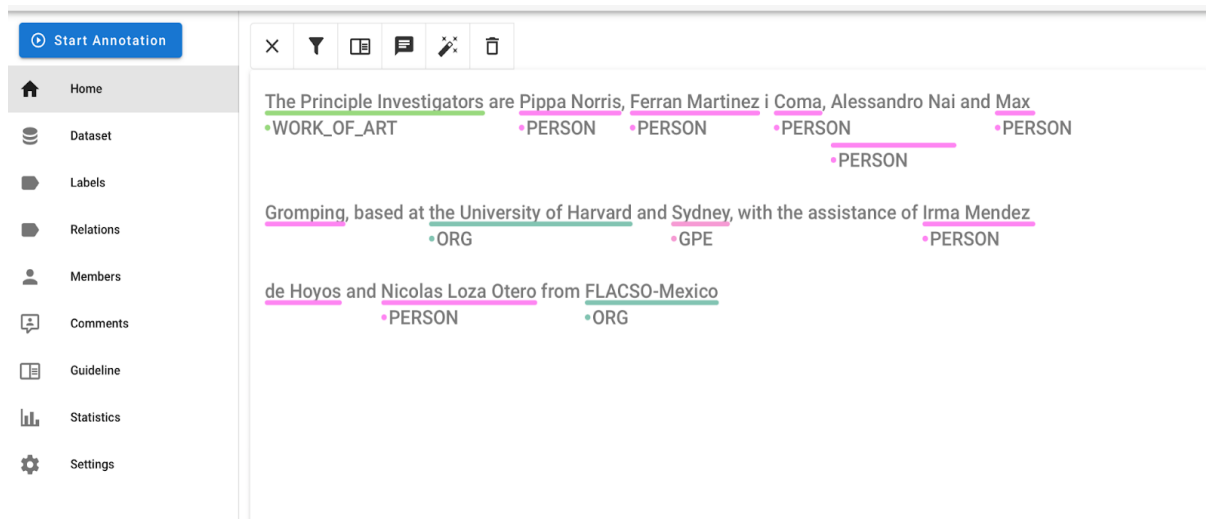


Figure 3. Automatic annotation in the Doccano text annotation tool

The utilisation of ML tools appears promising for automated metadata enrichment, but it is quite resource consuming if we wish to achieve the highest possible quality. It is manifest that only collaboration, additionally on an international level, will turn these explorations into standardised methods.

2.5 Beyond the CMDI use case

The CLARIAH+ project and the CMDI use case gave us the opportunity to explore and comprehend the first two parts of the Common Framework pipeline and helped us to demonstrate and test the complete pipeline: metadata extraction, metadata transformation and loading (archiving) of the enhanced metadata, in an instance of Dataverse.

Throughout this work we have taken the opportunity to collaborate with various other research groups and organisations:

- Extraction and evaluation of CMDI metadata fields, based on all CMDI metadata archived in DANS-EASY, in collaboration with the ODISSEI³⁷ project
- Definition of a core set of CMDI metadata fields in the cooperation with the CLARIN community³⁸
- Creation of a workflow for the prediction and linking of concepts from external controlled vocabularies to the CMDI metadata values (metadata enrichment), in collaboration with the CESSDA³⁹ community
- Extension of the *Common Framework* with the support for controlled vocabularies to create metadata that is available in a FAIR way, joining forces with Netwerk Digitaal Erfgoed⁴⁰ (NDE) team
- Extension of the export functionality of Dataverse to export deposited CMDI metadata back to the original CMDI format, together with KNAW Humanities Cluster (HuC)

The envisioned use of the *Common Framework* workflow as shown in explorations reported in this paper is two-fold: primarily, it informs CLARIAH+ about possibilities and challenges when it comes to the interoperability of metadata schemes; secondly, it informs DANS, as service provider of a long-term archive, about a portfolio of registered microservices which form a generic and extensible pipeline. DANS is currently migrating its research data archiving service from a Fedora-based platform (DANS-

³⁷ <https://odissei-data.nl/en/>

³⁸ Working document of mapping by Goosen, T., Broeder, D., Windhouwer, M., Köning, A., Labropoulou, P., Conzett, P., Van Uytvanck, D., Oleksy, M., Ohren, O.P., Tykhonov, V., De Vries, J., <https://docs.google.com/spreadsheets/d/1zKR5ErqL3wRX4tOL37110-34jXVP0gNzgU2vFsLrbcl/edit>

³⁹ <https://www.cessda.eu/>

⁴⁰ <https://netwerkdigitaal erfgoed.nl/>

EASY) to other platforms including introducing Dataverse to function as the DANS Data Stations as a specific repository service for designated communities (Wals, 2021). The designated communities will be served with larger metadata aggregations that include references to data not curated and/or hosted by DANS. The exploration described in this paper bases its analytic part on the current production system while, at the same time, informs the on-going migration process.

3 Future work

A substantial amount of work has been completed, but we are not finished yet. From the CMDI use case we discovered that CMDI as a standard is lacking a defined core set of CMDI metadata (Goosen et al., 2014). We remain in close cooperation with the CLARIN CMDI taskforce working on a proposal for, and acceptance of, a core set of CMDI metadata as a recommendation for all CLARIN centres.

The CMDI use case gave us the opportunity to prove the *Common Framework* approach. The following steps are to extend this Framework and to implement it for other cases. Beyond the extension of the Citation Core set of Dataverse, it is envisioned to support a link between other ‘indexing’ metadata fields to the other Knowledge Organisation Systems providers. In particular, we think here of recommended FAIR controlled vocabularies and ontologies which potentially may become part of the set of metadata fields (Wilkinson et al., 2016; Broeder et al., 2021; Wang, M. et al., 2021). Coming back to the CMDI case, this could lead to linkages of recognized, or any, CMDI metadata values to a recommended ontology or controlled vocabulary with the aim to produce ‘5-star Linked Open Data’⁴¹.

To contribute further to the FAIRification of controlled vocabularies and other KOS or Semantic artefacts we wish to experiment further with the creation of a semi-automatic workflow, using a Skosmos API, to query Skosmos representations of recommended controlled vocabularies. Therefore, explorations of the NDE’s Network of Terms⁴² GraphQL⁴³ endpoint will be continued to create links between appropriate controlled vocabularies for the terms extracted from the CMDI fields. These metadata fields will link to the CMDI component registry in the CMDI metadata schema.

Within the ODISSEI project DANS is going to work further on the creation of a production implementation of the microservices infrastructure. In the recently granted project FAIRCORE4EOSC⁴⁴ it is likely that DANS will be migrating the registries and brokering of microservices for schematic and semantic transformation/enhancement to the EOSC⁴⁵. DANS will continue to host some of the (micro)services, which ones are still a topic of debate. All future work of DANS will be shepherding transformations, enhancements, crosswalks, etc to a microservice/registry architecture.

All of our insights and workflows will be shared with the CLARIN and CLARIAH communities and we are looking for more collaborations on semantic mappings that could be used to get an appropriate ontology linkage not only on value level but also between fields available in CMDI Component Registry.

The *Common Framework* can help to support the enrichment of metadata, may aid the making of CLARIN datasets findable and accessible, and ultimately also supports Reusability. FAIR compliance automatic assessment tools, such as F-UJI⁴⁶, can be included in the *Common Framework* to evaluate the FAIRness of the metadata (Devaraju et al., 2020, 2021).

4 Conclusion

Our experimental work of building a *Common Framework* to expose CMDI metadata via a DANS discovery service relates to the migration of the DANS archive service to (a) newly to build DANS Data Station(s), which will serve as a basis for the discovery and OAI-PMH⁴⁷ harvesting services for the CLARIN researcher community and beyond.

⁴¹ https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

⁴² <https://github.com/netwerk-digitaal-erfgoed>

⁴³ <https://graphql.org/>

⁴⁴ <https://dans.knaw.nl/en/news/consortium-led-by-dans-acquires-a-major-european-grant-to-make-eosc-more-fair/>

⁴⁵ <https://ec.europa.eu/>

⁴⁶ <https://www.f-uji.net/>

⁴⁷ <http://www.openarchives.org/pmh/>

This paper describes the first two steps in our ETL-pipeline: 1) the extraction of the CMDI metadata from CMDI metadata files archived in DANS-EASY, and 2) the transformation of this metadata information. The third step of the pipeline, the loading of the transformed metadata into the new Data Station, is performed as a proof-of-concept in a separate Dataverse instance, one could say an envisioned Dataverse-based Data Station. The work is ongoing and the challenges we reflect upon, when addressed, are unavoidably leading to new challenges. For instance, we have been able to extend the Dataverse metadata model with a proposed core set of CMDI metadata which serves the needs of DANS as a basis for the discovery service. This resulted in a flexible solution which is easy to adjust in the event that the core set of CMDI metadata will be changed in the future. Its implementation in production services is still a challenge ahead.

To arrive at the proposed core set of CMDI metadata, we have analysed all linguistic and oral history datasets containing CMDI metadata stored in the DANS-EASY archive with the CMDI exploration tool. With the same tool we were able to transform each CMDI metadata file to the proposed core set.

To increase the FAIRness of the new metadata, we explored the possibilities of enriching the metadata with recommended external controlled vocabularies. This exploration has led to a flexible and generic solution to add custom external controlled vocabularies to Dataverse beyond the immediate CMDI use case. A semi-automatic workflow, which uses a Skosmos API, was developed to query any Skosmos representation of the recommended external controlled vocabularies. The NDE's Network of Terms GraphQL endpoint was used to make linkage to the appropriate controlled vocabularies for the terms extracted from the CMDI fields.

To extend the semi-automatic workflow we started to explore the possibilities of a semantic gateway. We started a proof-of-concept with a semantic gateway lookup API. This API is able to return a list of standardised concepts based on the selected vocabulary and a term. This will help to link each field in the proposed core set of metadata to the appropriate controlled vocabulary.

To complete the circle, we are currently in the phase of investigating the export of the Dataverse metadata back to the original CMDI format. The basic requirement for this should be that the Dataverse metadata schema must have CMDI metadata that can be extended with custom components, which are used by the different CLARIN centres. Secondly, the original relationships between fields and concepts should be kept whereby the custom components should be added to a SKOS schema. If this is possible, then we should be able to reproduce the original CMDI metadata, which could be offered for download to any user without losing the authority and provenance of the original metadata.

The basis of our work lies in reusing and exploring new techniques, (micro)services and the basic ideas behind the *Common Framework* are not only to solve long standing problems, but also to build flexible solutions for different communities. This is the main reason for the setup of a microservice oriented pipeline. Being part of different communities has helped us to create a broad support base amongst these communities. In the meantime, multiple communities, organisations and projects are testing and exploring our experimental work and connecting it to their own infrastructures, providing us with feedback to improve the microservices leading to sustainable infrastructure.

This work has taught us that looking to the future and setting ourselves some big challenges not only leads to innovative ideas and solutions, but it also leads to further new challenges. These challenges are motivating us to build sustainable solutions with and for the communities by exploring new technologies. These new technologies furthermore allow us to circumvent existing technology-lock-ins and they also demonstrate how, via microservices and a distributed approach, new methods of aligning and enriching metadata can be created.

Implementing these solutions in a sustainable infrastructure for long-term preservation archives is yet another challenge which we did not discuss in this paper. An important aspect when it comes to the implementation is the ownership of (meta)data and recommended controlled vocabularies, provenance and authorisation. We have demonstrated how technologies such as machine-learning approaches can be used to clean, enrich and harmonise metadata. We have also indicated communities must be involved in these technological developments, and how to implement to meet their needs. However, it remains to be seen and investigated how these technologies will be used in daily work by data producers and consumers (Borgman et al., 2019) and how they change the work of data managers and archivists. Moreover, more work is needed to investigate what are possible consequences for the certification process, and in general, which monitoring and governance policies are required for an envisioned network of distributed service providers that is required to remain stable over time.

Acknowledgement

This work has been funded by the CLARIAH Plus project, <https://www.clariah.nl/>, more particular in Work Packages 2, 3 and partly 4.

References

- Broeder, D., Budroni, P., Degl'Innocenti, E., Le Franc, Y., Hugo, W., Weiland, C., Wittenberg, P. and Zwolf, C. M. 2021. *SEMAF: A Proposal for a Flexible Semantic Mapping Framework*. Zenodo. <http://doi.org/10.5281/zenodo.4651421>
- Borgman, C., Scharnhorst, A., Golshan, M. S. 2019. Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, 70(8):888-904. DOI: 10.1002/asi.24172; preprint version: <https://arxiv.org/abs/1802.02689>
- Conzett, P., Goosen, T., Scharnhorst, A., Tykhonov, V., Van Uytvanck, D., de Vries, J. and Wittenberg, M. 2020. *How to weave domain specific information sources into a large, FAIR data fabric for the Digital Humanities? The use of the Dataverse platform*. Zenodo. <https://doi.org/10.5281/zenodo.3879031>
- Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., White, A. 2020. *FAIRsFAIR Data Object Assessment Metrics*. Working paper. Zenodo. <https://doi.org/10.5281/zenodo.4081213>
- Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Åkerman, V., L'Hours, H., Davidson, J., Diepenbroek, M. 2021. From Conceptualization to Implementation: FAIR Assessment of Research Data Objects. *Data Science Journal*, 20(1):4. <https://doi.org/10.5334/dsj-2021-004>
- Durco, M., Lorenzini, M., Sugimoto, G., 2018. Something will be connected - Semantic mapping from CMDI to Parthenos Entities. *Selected papers from the CLARIN Annual Conference 2017, Linköping Electronic Conference Proceedings*, 147(3): 25-35
- Durco, M., & Windhouwer, M. 2014. The CMD Cloud. In *Proceedings of LREC 2014: Ninth International Conference on Language Resources and Evaluation*. http://www.lrec-conf.org/proceedings/lrec2014/pdf/156_Paper.pdf
- European Commission, Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., Ojstersek, M., Choirat, C., van de Sanden, M., Coppens, F. 2021. *EOSC interoperability framework: report from the EOSC Executive Board Working Groups FAIR and Architecture*. Publications Office. <https://data.europa.eu/doi/10.2777/620649>
- Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Durco, M., Schonefeld, O. 2015. CMDI 1.2: Improvements in the CLARIN Component Metadata Infrastructure. *Selected Papers from the CLARIN Conference 2014, Linköping Electronic Conference Proceedings*, 116(004):36-53. <http://www.ep.liu.se/ecp/116/004/ecp15116004.pdf>
- Guéret, C., Chambers, T., Reijnhoudt, L., van der Most, F., Scharnhorst, A. 2013. *Genericity versus expressivity - an exercise in semantic interoperable research information systems for Web Science [Digital Libraries]*. Working paper. <http://arxiv.org/abs/1304.5743>
- Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., van Uytvanck, D. 2014. *CLARIN's Virtual Language Observatory (VLO) under scrutiny-The VLO taskforce of the CLARIN-D centres*. Working paper. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/3210/file/Haaf_Fankhauser_CLARINs_virtual_language_observatory_under_scrutiny_2014.pdf
- Hugo, W., Le Franc, Y., Coen, G., Parland-von Essen, J., Bonino, L. 2020, *D2.5 FAIR Semantics Recommendations Second Iteration*. Working paper. Zenodo. <https://doi.org/10.5281/zenodo.4314321>
- ISO 24622-1:2015. 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. 2019. *Language resource management – Component metadata infrastructure (CMDI) – Part 2: The Component metadata specification language*. Standard, International Organization for Standardization, Geneva, CH
- McIlwaine, I. C. 2010. Universal Decimal Classification (UDC), in M. J. Bates, M. N. Maack (eds.), *Encyclopedia of Library and Information Sciences, Third Edition (3rd ed.)*. CRC Press.1(1):5432-5439. <https://doi.org/10.1081/E-ELIS3-120043532>

- Myers, J., Tykhonov, V. 2021. *Proposal on the ontologies and external controlled vocabularies support in Dataverse*. Working paper. Zenodo. <https://doi.org/10.5281/zenodo.5845540>
- Odiijk, J. E. J. M. 2016. Linguistic research using CLARIN. *Lingua*, 178:1-4
- Shlomo, N. 2019. Overview of Data Linkage Methods for Policy Design and Evaluation, in Crato, N., Paruolo, P. (Eds.) *Data-Driven Policy Impact Evaluation*. Springer, Cham. https://doi.org/10.1007/978-3-319-78461-8_4
- Scharnhorst, A., Smiraglia, R. P. 2021. The Need for Knowledge Organization. Introduction to *Linking Knowledge: Linked Open Data for Knowledge Organization (Chapter 1)*. In R. P. Smiraglia; A. Scharnhorst (Eds.), *Linking Knowledge*. Ergon –Nomos, Baden-Baden, pp. 1-23. <https://doi.org/10.5771/9783956506611-1>
- Smiraglia, R. P., Scharnhorst, A., Akdag Salah, A., Gao, C. 2013. UDC in Action, in A. Slavic, A. Akdag Salah, & S. Davies (Eds.), *Classification and visualization: interfaces to knowledge*, pp. 259–270. Ergon Verlag, Würzburg. Preprint available at <http://arxiv.org/abs/1306.3783>
- Smiraglia, R.P., Scharnhorst, A. 2021. *Linking Knowledge. Linked Open Data for Knowledge Organization and Visualization*. Ergon-Nomos, Baden-Baden. <https://doi.org/10.5771/9783956506611>
- Svenonius, E. 2000. *The Intellectual Foundation of Information Organization*. The MIT Press, Cambridge, USA.
- Tennis, J. T. 2018. Intellectual history, history of ideas, and subject ontogeny. In *Challenges and Opportunities for Knowledge Organization in the Digital Age. Proceedings of the Fifteenth International ISKO Conference*. pp. 308 – 313. Ergon, Baden-Baden. <https://doi.org/10.5771/9783956504211-308>
- Tykhonov, V. 2021. *Controlled vocabularies and ontologies in Dataverse data repository*. Presentation at the *Dataverse Community Meeting 2021*. <https://doi.org/10.5281/zenodo.5838161>
- Tykhonov, V., de Vries, J., Scharnhorst, A., Admiraal, F., Indarto, E., Priddy, M. 2021. *Flexible metadata schemes for research data repositories*. Presentation at the *CLARIN Annual conference 2021*. Zenodo. <https://doi.org/10.5281/zenodo.5838156>
- Tykhonov, V., Scharnhorst, A. 2021. *Flexibility in Metadata Schemes and Standardisation: the Case of CMDI and DANS Research Data Repositories*. Presentation in the series *ISKO Knowledge Organisation Research Observatory*, November 24, 2021. Zenodo. <https://doi.org/10.5281/zenodo.5838109>
- Tykhonov, V. 2021. *CLARIN CMDI use case and flexible metadata schemes*. Presentation at the *CLARIAH interest group Linked Open Data*, 4 November 2021. <https://doi.org/10.5281/zenodo.5838132>
- Wals, H. 2021. *Focus on FAIR: DANS 2021-2025*. The Hague. Viewed 26 April 2022. https://dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/DANS-2021-2025/UK_DANS20212025.pdf
- Wang, M., Qiu, L., Wang, X. 2021. *A Survey on Knowledge Graph Embeddings for Link Prediction*. *Symmetry*, 13:458. <https://doi.org/10.3390/sym13030485>
- Wang, Y., Kadiyala, H., Rubin, J. 2021. Promises and challenges of microservices: an exploratory study. *Empirical Software Engineering* 26:63. <https://doi.org/10.1007/s10664-020-09910-y>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Nature*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Windhouwer, M., Broeder, D., & van Uytvanck, D. 2012. A CMD core model for CLARIN web services, in *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, pp. 41-48. <http://hdl.handle.net/11858/00-001M-0000-000F-A418-0>

Bagman – A Tool that Supports Researchers Archiving Their Data

Claus Zinn

Seminar für Sprachwissenschaft
Universität of Tübingen, Germany
claus.zinn@uni-tuebingen.de

Abstract

Getting researchers to archive their data properly is hard. Many factors are at play. In this paper, we present *Bagman*, a software that aims at alleviating research data management significantly. *Bagman* is a web-based software that supports researchers to package their data, assign a minimal set of metadata for their description, define a licence for the data's future distribution, and to submit the entire package in a safe manner to an archive of their choice.

1 Motivation

Research data management is an essential ingredient of good scientific practise. Theories explain the data, and for one researcher to validate another researcher's theoretic models, the inspection of data is central. Nevertheless, many researchers regard the management of research data as a necessary evil. Although one clearly acknowledges the benefits of proper research data management, it is also perceived as something that is not done with overwhelming desire or pleasure.

Fear of scientific scrutiny and competition aside, proper research data management feels like household chores; one needs to make an inventory of all research data, clean-up the data, iron-out a proper file and directory structure of all data, document the procedures and scripts for data annotation and analysis *etc.* When everything is in order, one needs to describe the data with metadata, and then bundle and safely transfer it to an archive of one's choice, so that eventually – once it is ingested into the archive and published – fellow researchers can find and make proper use of it.

The assignment of metadata is a particular nuisance. For this, researchers have to become familiar with metadata standards, registries, profiles, editors, validators, and best practises. Moreover, researchers are expected to take care of licensing issues, and last but not least, know about archives that are well suited to host their precious data.

Our new software, *Bagman*, aims at supporting researchers in all of the aforementioned areas to ease their pain as much as possible. At the same time, the *Bagman* developers strive to improve overall metadata quality, and also support archive managers to receive properly packaged research data.

2 Background

Getting your research data archived constitutes a workflow that varies across institutions. Details aside, it includes data packaging, metadata description, and transfer. Each step is accompanied by some quality control to minimize mishaps in these processes.

2.1 Packaging

In the worst case, researchers send their archive managers an email where all data is attached to the email. Sometimes data is put into some cloud space, or on portable storage devices for manual delivery. Such worst case scenarios often include data loss, files whose formats do not comply with archiving standards or whose names disobey naming conventions. Moreover, metadata descriptions might be anything from absent, incomplete or invalid XML. To avoid such mishaps, the art of packaging needs appreciation.

There are a number of tools that help researchers to bundle their research data into a single package. The open source software `docuteam packer` [URL-1] helps users bundling research data into a single package that can then be transferred to archives (Docuteam, 2018). The stand-alone Java application turns files into a *Submission Information Package* (SIP), a single data package that is delivered

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

to an archive or repository for (semi-)automatic ingestion, and which contains technical, structural and descriptive metadata in METS, PREMIS and EAD ([www.loc.gov/\[mets|premis|ead\]](http://www.loc.gov/[mets|premis|ead])).

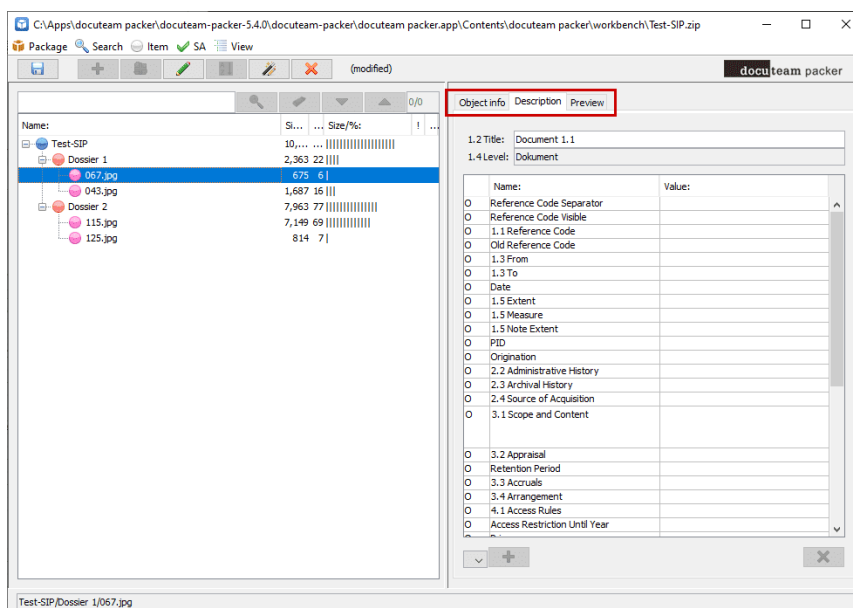


Figure 1: Software docuteam packer.

Fig. 1 depicts how users of Docuteam packer can add their research data in an incremental manner to the SIP. New files can be added to the file tree, and parts of the tree can be rearranged; also, for each object in the tree, metadata can be assigned. Usually, both researchers and archive managers will use the software. Researchers will use it to organize their research data into a tree, and to assign metadata to it to the best of their knowledge; then archive managers will use the software to complement metadata where it is missing.

```
myfirstbag/
├── manifest-md5.txt
│   ├── (49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png)
│   └── (408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt)
├── bagit.txt
│   ├── (BagIt-version: 1.0)
│   └── (Tag-File-Character-Encoding: UTF-8)
└── \--- data/
    ├── 27613-h/images/q172.png
    │   ├── (... image bytes ...)
    │   └── (...)
    ├── 27613-h/images/q172.txt
    │   ├── (... OCR text ...)
    │   └── (...)
    └── ....
```

Figure 2: A simple bag.

A software called `Bagger` was created for the U.S. Library of Congress as a tool [URL-2] to produce a package of data files according to the BagIt specification (Kunze et al., 2018). The specification is a set of hierarchical file layout conventions for storage and transfer of arbitrary digital content. Simply speaking, it can be seen as a shopping cart (bag) together with a shopping bill that lists each of the items with its location (path) and its price (an MD5 or SHA checksum). Those who receive the bag can use the inventory to check whether all goods were received in a complete and correct manner.

A simple example bag is given in Fig. 2. The figure shows the bag’s manifest file (the shopping bill) together with the inventory listed under data as well as some technical metadata about the BagIt version used and the file encoding.

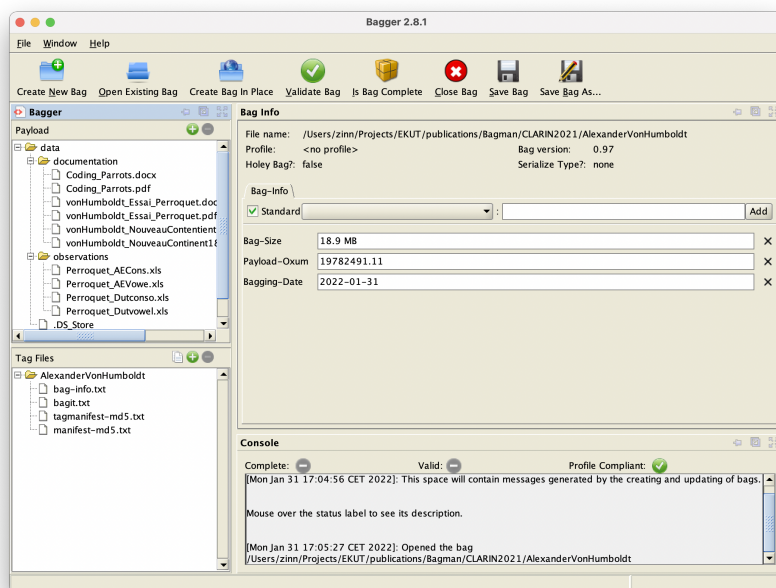


Figure 3: The Bagger Tool from the Library of Congress on our example data.

Fig. 3 shows the Bagger tool, a Java-based and desktop-bound application, in action. The functionality “Create Bag In Place” transforms a given data location on the user’s hard drive into a bag that is conform to the BagIt specification. That is, Bagger moves all research data into a subdirectory called data, computes the checksum for each file, and generates the file “manifest-md5.txt” along with the other tag files. In principle, users could be asked to install the Bagger application onto their local machine, create a bag, compress the bag into a zip archive, and then send it to their archive manager; the archive manager on the receiving end then unzips the archive and then uses the Bagger application to validate the bag.

Our software, Bagman, makes use of the BagIt format to help CLARIN researchers packing-up their research data so that it can be transferred to an archive in a correct and complete manner. Similar to docuteam packer, users are given the opportunity to describe their research data with metadata. Rather than asking users to fill out rather technical tables (see right-hand side of Fig. 1), Bagman aims to provide a more user-friendly approach by avoiding metadata jargon.

2.2 Metadata

Metadata plays a key role in any research infrastructure. Good metadata ensures that research data or any digital object are Findable, Accessible, Interoperable and Reproducible. The area of metadata research and practise is vast with many hundreds of metadata standards in use, and hundreds of policies in place to ensure that the FAIR principles are being followed [URL-3]. To support reproducible computational research, metadata formats must be sufficiently expressive to describe input (raw data, intermediate data), tools to process such data (with their version, dependencies, licence *etc.*), statistical reports and notebooks (*e.g.*, session variables, parameters), pipelines (dependencies between tools, provenance), and the resulting scientific publication (research domain, keywords, attribution *etc.*), see (Leipzig et al., 2021) for an overview.

In the CLARIN community, for researchers to assign metadata to data, they need to make use of the CMDI metadata framework (Broeder et al., 2012). For many researchers, this exercise feels like taming

multi-headed monsters in a landscape that feels rough and bracketed from every angle. Researchers need to consult the CMDI component registry [URL-4] to find a metadata profile that best fits their research data, and once they have identified a profile, they have to instantiate it to the best of their knowledge. This is not a trivial matter given that there are hundreds of profiles to choose from, but not a single metadata editor that gives intelligent help with instantiating the numerous different metadata fields.

No wonder, most CMDI-based descriptions have a rather poor descriptive power, taming the beast is exhaustive, and at some point one rather leaves it alone. As a result, researchers must be supported by dedicated archive management staff that is knowledgeable about the CMDI zoo of beasts, and that is armed with XML magic, best practises, and metadata processing tools to keep them at bay.

In Bagman, users are kept away from editing CMDI content directly. Information is gathered via simple forms, and information stemming from bagged resources is automatically added to the CMDI description. As a result, Bagman users are empowered to provide administrative, descriptive, and technical metadata with ease and minimal effort.

2.3 Archiving

The CLARIN infrastructure offers its community members a good number of repositories to store, preserve, and make available to others their research data. The CLARIN Virtual Language Observatory lists nearly 50 different data providers that host over 800 collections of valuable language-related resources. Finding the right archive for your research data is by means trivial when your home institution fails to provide an archive that fits your needs such as content fit or certification requirements, see [URL-8].

The German CLARIN website offers a “find your archive” service that helps researchers identifying the archive that is best suited to host their data [URL-5]. Users are requested to answer questions about the modality of their research data (spoken language, written language, multi-modal language, sign language), its lingual type (German, multi-lingual, historical *etc*), the type of their resource (*e.g.*, lexicon, corpus, treebank), and whether they choose a public licence or not. As a result, the centres that fit the answers best are returned, together with the contact details of the respective archive managers.

Bagman will use the information submitted by the user to suggest archives that are suitable for hosting the user’s research data. Once the user selected the archive, the bag will be safely transferred to a neutral place; the archive manager can download the bag from there, inspect the package, and then contact the user to proceed with the archiving procedure. Bagman hence aims at acting as a broker between researcher and archive manager.

3 Bagman

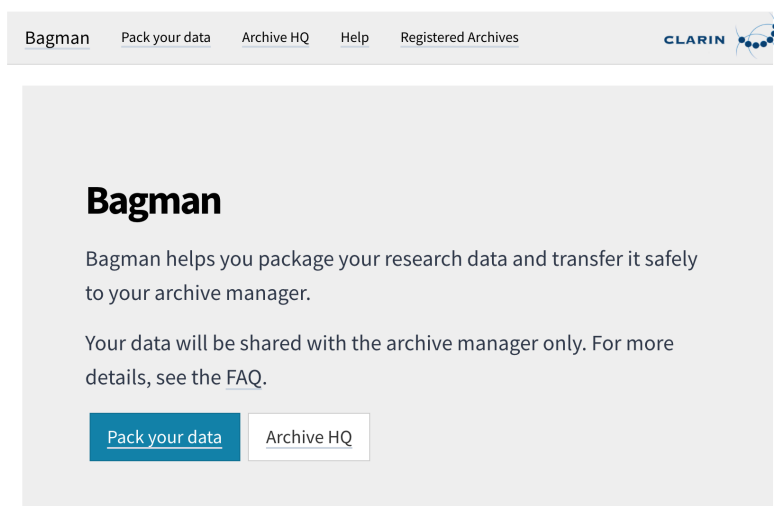


Figure 4: Bagman - Welcome Page.

Bagman aims at supporting researchers and archive managers alike. The software uses `Java` for the back-end and `react-js` for the front-end. Fig. 4 depicts the welcome page of Bagman; it gives access to its two core functionalities: “Pack your data” and “Archive HQ”. The first functionality is targeted at researchers who want to archive their research data; the second one is aimed at archive managers to get access to the research data packages submitted by users. In this paper, we will focus on the first aspect. Fig. 5 depicts Bagman’s user interface for collecting data from its users via simple forms.

Figure 5: Bagman - Requesting Metadata.

Researchers are requested to describe their research data with respect to the project where it has been collected and the researchers and their organisations that were involved. Users then classify their data in terms of a resource type and by answering a number of targeted follow-up questions about the chosen type. In the fifth step, users can select a licence for their research data. In the sixth step, users can upload their data by selecting a directory from their file system, see top-left part of Fig. 6. Note that some icons in the resulting tree are highlighted in red to signal file formats not suitable for archiving. Here, users are encouraged to convert, say, proprietary file formats to non-proprietary ones, or to delete superfluous ones. Note that Bagman delegates the main task for organising directory structures to users’ existing tools such as Finder (Mac OS), Explorer (Windows), or Files (Ubuntu), and file conversion software, say, Numbers, Excel, or OpenOffice. Once users have post-processed the directory tree, they can prepare the submission process (last step). Preparation includes the *automatic* generation of a CMDI file from known inputs as well as the submission package, the bag where all files all listed together with their checksums (see top-right and bottom part of Fig. 6). The back-end of Bagman takes care of all storage of research data, and it also implements basic functionality for CMDI generation. In detail, the back-end implements an API for (i) the generation of XML-based CMDI from JSON input, which is passed on from the client; (ii) the transferal of bags in ZIP format from client to server as well as methods for getting and deleting bags for archive management. Bagman also implements functionality for matching a bag with an archive that is best suited for hosting it.

Fig. 7 show a fragment of the CMDI file for the component `ResourceProxyList`, which is a

Uploaded research data

Note. Files with red icons use file formats unsuitable for archiving

Filter with:

- data
 - AlexanderVonHumboldt
 - observations
 - Perroquet_Dutvowel.xls
 - Perroquet_AECons.xls
 - Perroquet_AEVowe.xls
 - Perroquet_Dutconso.xls
 - documentation
 - Coding_Parrots.pdf
 - vonHumboldt_NouveauContentient1814.pdf
 - Coding_Parrots.docx
 - vonHumboldt_Essai_Perroquet.pdf
 - vonHumboldt_NouveauContinent1814.docx
 - vonHumboldt_Essai_Perroquet.docx

Bag Info

Label	Value
Source-Organization	Eberhard Karls Universität Tübingen
Contact-Name	Alexander von Humboldt
Contact-Phone	+49 (0) 7071-29 73968
Contact-Email	avh@uni-tuebingen.de
Description	Second Language Acquisition in Parrots
Bagging-Date	2021-04-27
BagIt-Version:	1.0
Tag-File-Character-Encoding:	UTF-8
Bag-Count:	10
Bag-Size:	18.9 MB

I accept the terms and conditions (link follows...).

Bag Entries

File	Size	Mimetype	SHA256
AlexanderVonHumboldt/observations/Perroquet_AEVowe.xls	4835840	application/vnd.ms-excel	c4c9aa805fda4eea2dc1aed77638520427290f6bdd8d57d2ab3c2a99ce7c7c9a
AlexanderVonHumboldt/documentation/Coding_Parrots.pdf	55561	application/pdf	446c8f51286e25015270ff054beeafa68ab9fb8be537acb96f0f30c29dec0819
AlexanderVonHumboldt/documentation/vonHumboldt_NouveauContentient1814.pdf	99506	application/pdf	1c991c1f2599ebcc5ba82927b7382e64f4027b0f91cd8a450c5fc609a5c6e3c2
AlexanderVonHumboldt/documentation/Coding_Parrots.docx	14822	application/vnd.openxmlformats-officedocument.wordprocessingml.document	9dff7c8802debb5315301d46513a5dd6e207c38c14e09ae794f2f4fb0fa85518
AlexanderVonHumboldt/documentation/vonHumboldt_Essai_Perroquet.pdf	133324	application/pdf	7007172cf4807c11d17ae7a6bb204850d69ebdfb50094a2f7574724bc123f7e4

Figure 6: Bagman - Various Screenshots.

central part of the CMDI header. Each resource that our example user has uploaded is tagged as ResourceType “Resource” together with the mimetype that Bagman identified. The id attribute of the ResourceProxy component assigns a unique id to each resource.

Fig. 8 shows the corresponding ResourceProxyListInfo component, which reuses the aforementioned unique identifiers. Here, additional information about each resource is given, in particular, the resource’s file size and its checksum in the cryptographic encodings “md5”, “sha1”, and “sha256”.

4 Current State and Future Work

We have built a prototype of Bagman that implements its core functionality and which is now open for beta testing at the website <https://weblicht.sfs.uni-tuebingen.de/bagman/>. We invite readers to explore the tool and encourage their feedback. At the time of writing, only a single archive has been connected to Bagman to test and validate the transfer of data between researchers and archive managers. With research data temporarily stored on Bagman’s back-end, adding new archives to Bagman means giving their managers a login so that they can get access to the bags submitted to them. At the time of writing, Bagman supports the five major resource types hosted by TALAR, the Tübingen Archive of Language Resources [URL-6]; our software hence allows the automatic instantiation of CMDI profiles for the description of lexical resources (LexicalResourceProfile), text corpora (TextCorpusProfile), speech corpora (SpeechCorpusProfile), tools (ToolProfile), and experiments (ExperimentProfile), all identifiable via the Group Name “NaLiDa” in the CMDI component registry. The use of these profiles ensures that the corresponding resources can be easily found using faceted browsing in the Virtual Language Observatory, say, by searching the facets for language, collection, resource type, modality, or availability [URL-7].

The design of Bagman walks a fine line between researchers (often taking research data management

```

<cmd:Resources>
  <cmd:ResourceProxyList>
    <cmd:ResourceProxy id="id-2cf274a2-bed0-4737-86a1-25d353784b68">
      <cmd:ResourceType mimetype="application/vnd.ms-excel">Resource</cmd:ResourceType>
      <cmd:ResourceRef>observations/Perroquet_Dutvowel.xls</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-77030d3c-983f-4394-a5ed-6d4e3202e2fb">
      <cmd:ResourceType mimetype="application/vnd.ms-excel">Resource</cmd:ResourceType>
      <cmd:ResourceRef>observations/Perroquet_Dutconso.xls</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-117f8cfd-4b16-4e51-936f-6075f709d359">
      <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
      <cmd:ResourceRef>documentation/Coding_Parrots.pdf</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-dfb55091-87c8-4b69-9b63-fe299e80b18e">
      <cmd:ResourceType mimetype="application/vnd.ms-excel">Resource</cmd:ResourceType>
      <cmd:ResourceRef>observations/Perroquet_AECons.xls</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-c32cb84e-a621-4391-aba0-1fbedce415f1">
      <cmd:ResourceType mimetype="application/vnd.ms-excel">Resource</cmd:ResourceType>
      <cmd:ResourceRef>observations/Perroquet_AEVowe.xls</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-a8e9053a-11f4-437e-a240-e3ca7348269d">
      <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
      <cmd:ResourceRef>documentation/vonHumboldt_NouveauContentient1814.pdf</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-c06a7a20-31eb-42b9-8f2d-39ef133485bb">
      <cmd:ResourceType mimetype="application/vnd.openxmlformats-officedocument.wordprocessingml.document">Resource</cmd:ResourceType>
      <cmd:ResourceRef>documentation/Coding_Parrots.docx</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-3c188547-668c-4e90-836b-69c2f86fb382">
      <cmd:ResourceType mimetype="application/vnd.openxmlformats-officedocument.wordprocessingml.document">Resource</cmd:ResourceType>
      <cmd:ResourceRef>documentation/vonHumboldt_Essai_Perroquet.docx</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-04c27ac0-08ae-4607-b908-57ace1e203de">
      <cmd:ResourceType mimetype="application/vnd.openxmlformats-officedocument.wordprocessingml.document">Resource</cmd:ResourceType>
      <cmd:ResourceRef>documentation/vonHumboldt_NouveauContentient1814.docx</cmd:ResourceRef>
    </cmd:ResourceProxy><cmd:ResourceProxy id="id-a60ca8ef-8cc6-469a-9b95-b0a0188c9b4b">
      <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
      <cmd:ResourceRef>documentation/vonHumboldt_Essai_Perroquet.pdf</cmd:ResourceRef>
    </cmd:ResourceProxy></cmd:ResourceProxyList>
  <cmd:JournalFileProxyList> [4 lines]
  <cmd:ResourceRelationList> [1 line]
</cmd:Resources>

```

Figure 7: CMDI Excerpt – ResourceProxyList component.

as a necessary evil) and archive managers (taking it for something absolutely necessary, with an emphasis on “the more metadata the better”). When Bagman users, for instance, identify their data as a lexical resource, they are given the opportunity to specify the type of the lexicon (*e.g.*, dictionary, glossary, thesaurus), the type of the headword, and the subject language, but they may skip the step if they want to. Also, they can put more information about their resource in an open-ended lexicon description field when they feel that more information needs to be put somewhere. Note, however, that Bagman delegates any metadata-related issues to a subsequent one-to-one communication between researcher and archive manager. Metadata fields left open during a Bagman session can often be filled at a later stage when archive managers feel they require more information than researchers provided.

Bagman is browser-based software, and hence, special care needs to be taken to ensure that users can provide their input in a flexible, piece-wise manner. At any time, users can save the current session, that is, write-out all metadata that has been entered to their file system. At a later time, when users like to resume their work, they can then easily restore their session.

At the time of writing, Bagman is only connected to TALAR, but it supports all the archive’s profiles. For TALAR users, Bagman has entered production mode. The feedback we obtain from these real-world users informs the further development of Bagman, strengthening its usability and stability. Once Bagman has matured, we will ask other archives whether they want to be connected to Bagman, and we will investigate how their archiving requirements can be met with the software. Currently, it is too soon to speculate about the detailed implementation roadmap for the archiving aspect of Bagman. It is clear that other archives will like to see their metadata profiles and archiving policies supported. Here, Bagman would need to adapt its front-end to collect information specific to the new profiles, and the back-end to generate ready-to-use and valid CMDI instances that other archives are happy to work with.¹

Bagman does not prescribe any guidelines on the granularity of the research data that needs to be archived. Each set of resources is different, and Bagman *per se* does not attempt to promote a *one-size-fits-all model*.² Naturally, there exist research data that are not easily or adequately described with

¹At the time of writing, Bagman pre-fills some form fields such the organisation’s name or address. The default values of such fields are specific for TALAR users, but can be overwritten. With each new archive being connected, Bagman would also need to adapt its front-end to provide default values specific to the archive.

²As a rule of thumb, all research data created to support a scientific finding should be bundled into a single archival unit.

```

<ResourceProxyListInfo cmd:ComponentId="clarin.eu:cr1:c_1470820607607">
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673"
    cmd:ref="id-2cf274a2-bed0-4737-86a1-25d353784b68">
    <ResProxItemName>Perroquet_Dutvowel.xls</ResProxItemName>
    <ResProxFileName>observations/Perroquet_Dutvowel.xls</ResProxFileName>
    <FileSize>4866560</FileSize>
    <Checksums>
      <md5>e0288dce9a55ffdef4af1e73e650c747</md5>
      <sha1>636a3e8803a09c958994201c3fff1fff5879366dc</sha1>
      <sha256>23f35f38af09f1868e327b978d942dac6d5deaa345690881cfc60be59bf81267</sha256>
    </Checksums>
  </ResourceProxyInfo>
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673"
    cmd:ref="id-77030d3c-983f-4394-a5ed-6d4e3202e2fb">
    <ResProxItemName>Perroquet_Dutconso.xls</ResProxItemName>
    <ResProxFileName>observations/Perroquet_Dutconso.xls</ResProxFileName>
    <FileSize>4871680</FileSize>
    <Checksums>
      <md5>c614c88a495920aebecd29594074be40</md5>
      <sha1>aecbe906280a8b077cdc1a5ebc06c7393122e195</sha1>
      <sha256>7207f1cdf331e8435ebcb72271addacd5c83ae25ca1b40a5f8048a9ff4c08413</sha256>
    </Checksums>
  </ResourceProxyInfo>
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
  <ResourceProxyInfo cmd:ComponentId="clarin.eu:cr1:c_1361876010673" [10 lines]
</ResourceProxyListInfo>

```

Figure 8: CMDI Excerpt – ResourceProxyListInfo component.

Bagman and the CMDI profiles it currently supports. While the TALAR-based CMDI profiles have a good descriptive power, they might have shortcomings when it comes to the description of heterogeneous research data that researchers see as a single archival unit. For now, most users are unaware of Bagman. They contact the TALAR archive manager because they would like to have their resources deposited. Once the contact has been established and any open questions between the two parties addressed (e.g., granularity or licence issues), the users are then explicitly directed to Bagman to build, describe, and submit their package to the archive via Bagman.

One important aspect to Bagman’s usability is the packaging. When the archive manager is informed of a new bag being submitted via Bagman, he can download the bag from Bagman’s “Archive HQ” GUI, unzip the bag and run BagIt software to verify that the package has been transferred in a complete and correct manner.³ Our TALAR archive managers find this functionality very useful and reassuring indeed, and a necessary first step before looking into the CMDI, and contacting the researchers for any follow-ups, such as resolving metadata issues, or the drafting and signing of data depositing agreements.

Note that the use of the BagIt specification duplicates information that is also present in the CMDI file generated by Bagman, in particular, the information shown in Fig. 7 and Fig. 8. The duplication of such technical metadata, however, is well justified. The bag delivered to the archive is used to ensure that all research data is being transferred in a complete and correct manner. and archive managers can use the aforementioned toolchain to validate the bag. The CMDI file, of course, is used by metadata harvesters such as the VLO to being able to link to the resources the metadata describes.

Getting users to archive their research data is hard. Bagman offers users a single pit-stop approach to get their data archived without too much hassle. Bagman helps users with metadata management as it generates a CMDI automatically from the information and research data supplied by the user. Given such

³The command `python3 -m bagit --validate bag` verifies the bag.

data, Bagman then helps users to decide on an archive to host their resource, and then helps ensuring that all data is transferred to the archive in a complete and correct manner. In sum, Bagman fills-in a gap in the CLARIN infrastructure; its ease of use encourages users to get their data archived; and its automatic generation of CMDI from known inputs ensures the generation of expressive and high-quality metadata.

Acknowledgements

Our work was funded by the German Federal Ministry of Education and Research, the German Science Foundation (SFB-833), and CLARIN-D.

References

Broeder, D., Windhouwer, M., Van Uytvanck, D., Goosen, T., and Trippel, T. 2012. CMDI: a Component Meta-Data Infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR. Workshop at LREC-2012*, Istanbul, Turkey.

Docuteam. 2018. Software – our tools for digital archives. Available at <https://www.docuteam.ch/en/products/it-for-archives/software/>.

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

Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K., and Greenberg, J. 2021. The role of metadata in reproducible computational research. *Patterns*, 2(9). <https://doi.org/10.1016/j.patter.2021.100322>.

Links

[URL-1] Docuteam packer, see <https://docs.docuteam.ch/packer/6.1/en/index>.

[URL-2] Bagger, see <https://docs.docuteam.ch/packer/6.1/en/index>.

[URL-3] The FAIR principles, see <https://fairsharing.org>.

[URL-4] Component registry, see <https://catalog.clarin.eu/ds/ComponentRegistry>.

[URL-5] Centre finder, see <https://www.clarin-d.net/en/preparation/find-a-clarin-centre>.

[URL-6] TALAR, see <https://talar.sfb833.uni-tuebingen.de>.

[URL-7] Virtual Language Observatory, see <https://vlo.clarin.eu>.

[URL-8] Core Trust Seal, see <https://www.coretrustseal.org>.

ARCHE Suite: A Flexible Approach to Repository Metadata Management

Mateusz Żółtak
ACDH-CH OEAW
Vienna, Austria

`mateusz.zoltak@oeaw.ac.at`

Martina Trognitz
`martina.trognitz@oeaw.ac.at`

Matej Ďurčo
`matej.durco@oeaw.ac.at`

Abstract

This article presents an innovative approach to metadata handling implemented in the ARCHE Suite repository solution. It first discusses the technical requirements for metadata management and contrasts them with the shortcomings of existing solutions. Then, it demonstrates how the ARCHE Suite addresses those problems. After one year of use, we can assert that the approach implemented in the ARCHE Suite is viable and provides important benefits. We aim to establish the ARCHE Suite as an open-source repository solution to be used also by other parties.

1 Introduction

The Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) at the Austrian Academy of Sciences in Vienna runs the repository ARCHE for persistent hosting of humanities research data. ARCHE is certified as a CLARIN B-centre. Between 2017 and 2020, the underlying software technology we used was Fedora Commons version 4 with Blazegraph as a metadata store. Due to many serious shortcomings related to metadata management, the increasing amount of technical issues, and a lack of adequate alternatives, we decided to develop our own repository solution: the *ARCHE Suite*.

This paper specifies core requirements for metadata management and explains why they are not met by the existing repository solutions Fedora Commons (The Fedora Leadership Group, 2016), DSpace (Smith et al., 2013), Dataverse (King, 2007) or Invenio (Holm Nielsen, 2019)¹. We describe how the desired features have been implemented in our solution and how they are used in our metadata management workflows. Finally, we discuss the challenges posed by our solution and summarise our first-year experiences of using it.

2 Technical Requirements for Metadata Handling

Metadata is a vital part of every data repository, indispensable for finding, understanding, and reusing the data. To fully comply with the FAIR Data Principles that emphasise machine-actionability (Wilkinson et al., 2016), data and metadata have to be machine-readable and interoperable, which poses many challenges. The most important one includes ensuring metadata interoperability and consistency while preserving its descriptive precision. Handling these challenges governs our core technical requirements for the ARCHE Suite.

2.1 Ensuring Metadata Interoperability

In the humanities and cultural heritage disciplines, the tremendous amount of metadata standards (e.g., (Riley, 2010)) stands in the way of metadata interoperability. To overcome this, CLARIN has introduced the Component Metadata Infrastructure (CMDI) (Broeder et al., 2012), a standardised (ISO 24622-1, 24622-2) metadata framework with a built-in interoperability mechanism. Another compromise widely used across all disciplines is to apply the DCMI Metadata Terms (DCMI Usage Board, 2020), with the

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

¹Webpage for Fedora Commons: duraspace.org/fedora/, for DSpace: duraspace.org/dspace/, for Dataverse: dataverse.org/, and for Invenio: invenio-software.org/

caveat of losing the potentially richer metadata to one basic common set of metadata descriptors. The repository solutions most popular among CLARIN B centres - Fedora Commons 3 and DSpace (see Tables 1 and 2) - force users to use Dublin Core (DC) as a repository-native metadata format in a more or less explicit way.

CLARIN B Centre	City	Software
ASV Leipzig	Leipzig	Fedora Commons, v3
ACDH-CH - ARCHE	Vienna	ARCHE Suite
Bayerisches Archiv für Sprachsignale	Munich	own solution
Berlin-Brandenburg Academy of Sciences and Humanities	Berlin	Fedora Commons, v3
Center of Estonian Language Resources	Tartu	META-SHARE & own (Entu)
CLARIN.SI Language Technology Centre	Ljubljana	DSpace
Eberhard Karls Universität Tübingen	Tübingen	Fedora Commons, v3 (v4 planned)
Hamburger Zentrum für Sprachkorpora	Hamburg	Fedora Commons, v3
Institut für Maschinelle Sprachverarbeitung	Stuttgart	Fedora Commons
Instituut voor de Nederlandse Taal	Leiden	own
Leibniz-Institut für Deutsche Sprache	Mannheim	Fedora Commons
LINDAT/CLARIAH-CZ	Praha	DSpace
MPI for Psycholinguistics	Nijmegen	Fedora Commons, v3 (v4 planned)
PORTULAN CLARIN	Lisboa	META-SHARE & own
Språkbanken	Gothenburg	DSpace & own (Korp, etc.)
The ILC4CLARIN Centre at the Institute for Computational Linguistics	Pisa	DSpace
The Language Bank of Finland	Helsinki	META-SHARE & own (tools)
Universität des Saarlandes	Saarbrücken	Fedora Commons
ZIM Centre for Information Modelling	Graz	Fedora Commons, v3 (v4 planned)
CLARIN-PL Language Technology Centre	Wrocław	DSpace
CLARINO Bergen Center	Bergen	DSpace
CMU-TalkBank	Pittsburgh	own (talkbank)
The CLARIN Centre at the University of Copenhagen	Copenhagen	DSpace & eSciDoc

Table 1: Repository software solutions used by CLARIN B centres according to the CLARIN's Centre Registry. The information is based on a centre's registry entry or its latest CoreTrustSeal document.

In the last years, a new concept for (meta-)data interoperability has gained prominence: the Linked Open Data (LOD) principles with five levels (stars) of compliance (Berners-Lee, 2009). Four-star LOD (Berners-Lee, 2009; Holborn, 2014) requires data to be provided in a W3C-compliant standard like RDF (W3C et al., 2014) or SPARQL (W3C et al., 2013). This is easily met by using DC because of a well-defined mapping to RDF (W3C et al., 2014; Nilsson et al., 2008). The real challenge, however, is to additionally meet the requirements of five-star LOD, which includes the use of external links (Berners-Lee, 2009; Holborn, 2014). Using external URLs as DC term values meets the requirements but results in a repository inaccessible to human users, who expect human-readable labels like *'Karl Baedeker'* rather than URIs or URLs like *'arche.acdh.oeaw.ac.at/api/35998'*. Using both URLs and human-friendly text labels as values results in problems with DC properties used multiple times (e.g. *dc:creator*) because the corresponding labels and URLs cannot be paired anymore. Overall, the only viable solution to fully adopt five-star LOD seems to be providing full RDF support.

Fedora Commons 3 does not have any RDF support. This has been changed in Fedora Commons 4 where RDF became a native metadata format. Unfortunately, Fedora Commons 4 and 5 suffer from a serious feature drop compared to the previous version (most notably the lack of a search API and dissem-

Repository software	No. of centres
ARCHE Suite	1
DSpace	7
Entu	1
eSciDoc	1
Fedora Commons	9 (v3: 6)
META-SHARE	2
own solution	6

Table 2: Popularity of repository software solutions used by CLARIN B Centres listed in Table 1.

ination methods). As a result, the adoption of Fedora Commons 4 and 5 has never become widespread. The lack of the search API has been addressed in version 6² but the introduced API has no RDF support. On top of that, Fedora Commons 4-6 enforce a hard-coded metadata schema for all metadata properties managed by the service (media type, binary content size, creation and modification date, etc.).

Dataverse presents a mixed approach. On the input side, it requires metadata to follow a bespoke Dataverse schema making it interoperable with other Dataverse repositories only. On the output side, metadata can be serialised into a few schemas (Institute for Quantitative Social Science, 2021), e.g. schema.org’s Dataset RDF schema serialised as JSON-LD.

Invenio allows any metadata schema which can be defined using the JSON Schema (CERN et al., 2021). Such a solution can be considered RDF-compliant to a large extent because RDF metadata can be serialised as JSON-LD, and the resulting JSON-LD structure can be described in the JSON Schema. The limitation here is that there can be many valid RDF to JSON-LD serialisations, and it can be impossible to describe all of them using the JSON Schema.

DSpace defaults to Dublin Core but can be set up to accept any flat metadata schema. The limitation is that it requires metadata to be provided serialised as XML in the way that a property is represented as an XML tag and the value is the tag’s content. This is incompatible with RDF in two ways: First, it forbids ingestion of RDF metadata containing URI values because in RDF-XML the URIs are stored as XML tag attributes and not as a tag’s content. Second, the flat internal metadata model makes it impossible to store multiple values (URLs and labels) of the same metadata property in such a way that relationships between them (e.g. *this is a label for this URL*) are maintained. Despite the limitations on the data input side, DSpace allows a few RDF serialisation options on the output side. E.g. generation of an OAI-PMH record in the RDF-XML format with an XSLT stylesheet. Another option is to couple DSpace with a triplestore (DuraSpace, 2021).

The interoperability imperative combined with the heterogeneous formats landscape implies that most repositories have to handle more than one metadata format. The enforcement of a metadata schema (often DC) by a repository software is undesired as it either prevents the handling of domain-specific metadata schemas or requires extensive customisation. A typical way of overcoming limitations imposed on the metadata schema by a repository software is to materialise domain-specific formats as separate repository data streams. The main disadvantage of this approach is making the information redundant, which brings the risk of inconsistency. Furthermore, if a presentation format has to be changed, e.g. because a CMDI profile definition is updated, all materialised metadata records have to be regenerated and updated even if there is no change in the metadata values themselves. Similarly, if a metadata value changes, both the repository-native metadata format as well as all materialised metadata data streams have to be updated.

A better solution is to keep a single copy of all metadata values in a schema-agnostic metadata store and to allow for on-demand conversion to the desired metadata format with a templating system. DSpace and Fedora Commons have no embedded support for on-the-fly metadata conversion, Dataverse provides a fixed set of built-in conversions as described above, and Invenio allows to write custom metadata schema conversion plugins in Python.

²Fedora Commons 6 was released on 30th June 2021, while development on the ARCHE Suite had already begun by end of 2019. The information on Fedora Commons 6 provided here originate from its technical documentation and not from testing.

2.2 Ensuring Metadata Consistency

Ensuring metadata consistency, preferably at the ingest stage already, involves several aspects to be considered in the context of the repository management software. First, the way in which metadata checks are defined. This can be done either by specifying the allowed schema when using configuration files or by plugging in own code which performs the checks. Dataverse only supports the former method, Fedora Commons 4 only the latter, DSpace and Invenio both, and Fedora Commons 3 has no support for custom metadata checks. Executing a pluggable code only after the data were stored in the repository, like in Fedora Commons 4, does not allow for reliable metadata checks because it either allows the metadata to stay in an inconsistent state or rejects it without notifying the client about the ingestion failure.

The second aspect regarding metadata consistency concerns the software layer, in which the metadata restrictions are verified. To ensure that checks can not be bypassed, they have to be enforced by a single software component responsible for handling all data irrespective of the ingestion interface.

The third important factor is the ability to ingest the data using ACID — atomicity, consistency, isolation, durability — (Haerder and Reuter, 1983) transactions. It is especially important from the LOD perspective where consistency of one repository resource metadata may depend on a successful creation (or update) of another repository resource. Unfortunately, ACID transactions are poorly supported by existing repository solutions.

Invenio provides only a basic optimistic concurrency control on a single resource modification request level. Dataverse, DSpace and Fedora Commons 3 lack any concurrency control on the client API level and our experience with the previous ACDH-CH repository based on Fedora Commons 4 proved its transaction support to be intrinsically broken. Reasons for this are that Fedora Commons 4 and 5 lack a built-in search feature and the synchronisation with an external search engine like Solr or a triplestore is done only after the transaction commit. This makes it impossible for the ingestion client to search for any ingested data until the transaction's end. Furthermore, there is no locking system preventing parallel transactions from modifying the same repository resource (a lack of a so-called *transaction separation*). As a consequence, Fedora Commons 4 and 5 commit and rollback transaction operations provide no guarantee regarding the final state of resources modified by a transaction. Additionally, there are smaller issues like requests made within a transaction not extending the transaction timeout. The latter can lead to the failure of a large resource upload (e.g. few gigabytes in size) when the upload takes more time than the transaction timeout. The Fedora Commons 6 documentation suggests no changes in this regard.

2.3 Requirements List

To sum up, the desired repository solution should:

- Provide RDF support as the only viable way of fulfilling the five-star LOD principles
- Not enforce any particular metadata schema
- Avoid metadata duplication that comes from materialising metadata in different formats
- Allow for defining upon-ingestion metadata consistency checks in a flexible way
- Ensure metadata consistency in a way that cannot be easily bypassed
- Provide fully ACID transactions
- Allow for writing extensions in many programming languages.

Unfortunately, none of the existing solutions provides support for all the points from this list. For this reason, we have developed a new repository software: the ARCHE Suite.

3 The ARCHE Suite

The ARCHE Suite is a bespoke, in-house repository solution that we developed from scratch within half a year in 2020, including the migration from the old Fedora 4-based repository. Before going into production it underwent an external code review. The ARCHE Suite is built in a modular, service-oriented manner, consisting of multiple interconnected components that communicate through well-defined APIs (see Figure 1). All software components are available on GitHub³ and the documentation is provided at acdh-oeaw.github.io/arche-docs/.

³github.com/acdh-oeaw?q=arche

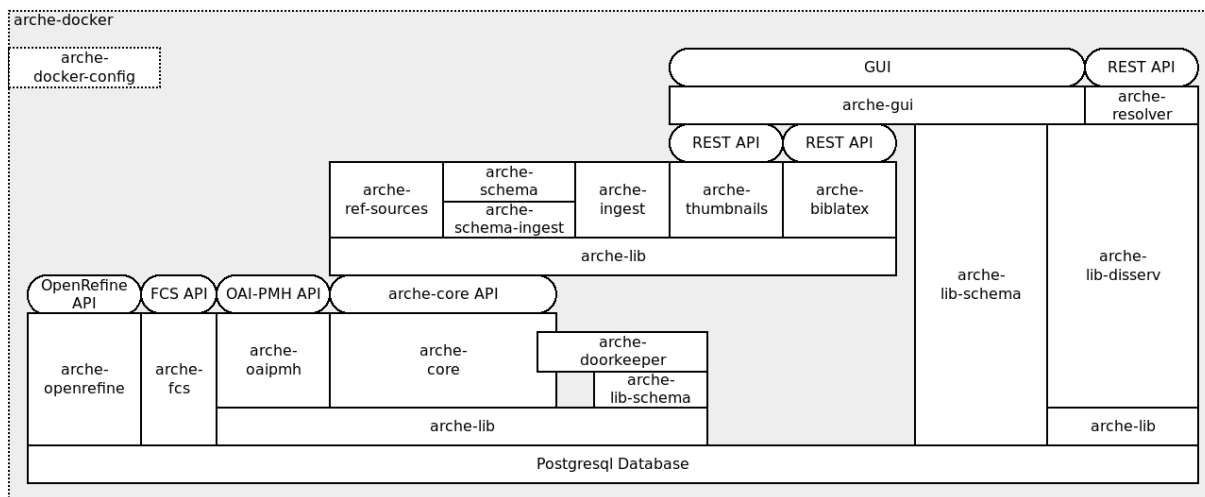


Figure 1: ARCHE Suite components. As can be seen, a microservice-based approach has been used.

Here, we detail the technical implementation of the metadata-related requirements formulated above. We focus on the developed software solution, ARCHE Suite, as opposed to ARCHE, the specific repository instance certified as a CLARIN B Centre provided by the ACDH-CH with the ARCHE Suite as the underlying technology. While ARCHE Suite is schema-agnostic, in ARCHE every resource must be described with metadata respecting a bespoke and elaborate schema (Trognitz and Ďurčo, 2018).

3.1 RDF Support

We decided to avoid dependency on a triplestore and to use a relational database as a metadata store instead. The database schema is developed in a way it can store any RDF data, i.e. does not enforce any particular RDF schema. There were two main reasons for this decision. First, using a triplestore makes it difficult to implement ACID transactions because triplestores do not recognize this concept. Second, using a relational database backend allowed us to significantly lower CPU and memory consumption of the repository (see Figure 2). On average we achieved 10 times lower memory usage and 10 to 25 times lower CPU usage. It is also important that we avoided resource usage peaks coming from the triplestore (see the middle of the right-hand column charts in Figure 2). Last but not least it sped up data ingestion by a factor of four. As a result, the ARCHE Suite supports RDF as metadata format both on the input and output side but does not natively provide a SPARQL endpoint. A dedicated search API is used instead. However, a triplestore can be paired with the ARCHE Suite either by using the plugins system described below or by periodic synchronisation. We already successfully tested the periodic synchronisation scenario.⁴

To compensate for the lack of a native SPARQL endpoint, the ARCHE Suite REST API allows to flexibly define the amount of linked data to be provided, e.g. it is possible to extend a REST API call response with metadata of *'all resources that are pointed to by a given resource'* or metadata of *'all resources that point to a given resource'* or metadata of *'all resources which can be reached by following a given RDF property'* or all of them. This solution proved to be very convenient and for performance reasons we strongly prefer it over a triplestore (see Figure 2)⁵.

The data model assumes a direct connection between the metadata RDF graph and the repository structure: Every node in an ingested RDF metadata graph corresponds to a repository resource. The repository can be configured either to automatically create repository resources when an unknown RDF graph node is found in the metadata graph or to treat it as a metadata inconsistency and raise an error.

Figure 3 illustrates this connection by showing how ingested RDF nodes are processed into repository resources. In the upper part of Figure 3 an RDF graph representing a collection with the title *Collection 1* and the author *John Doe*, who comes with one custom (*https://myNmsp/Doe/John*) and one external

⁴See the arche2sparql Docker image: github.com/csae8092/arche2sparql

⁵See also ARCHE REST API scalability testing on acdh-oeaw.github.io/arche-docs/aux/metadata_api_performance.html

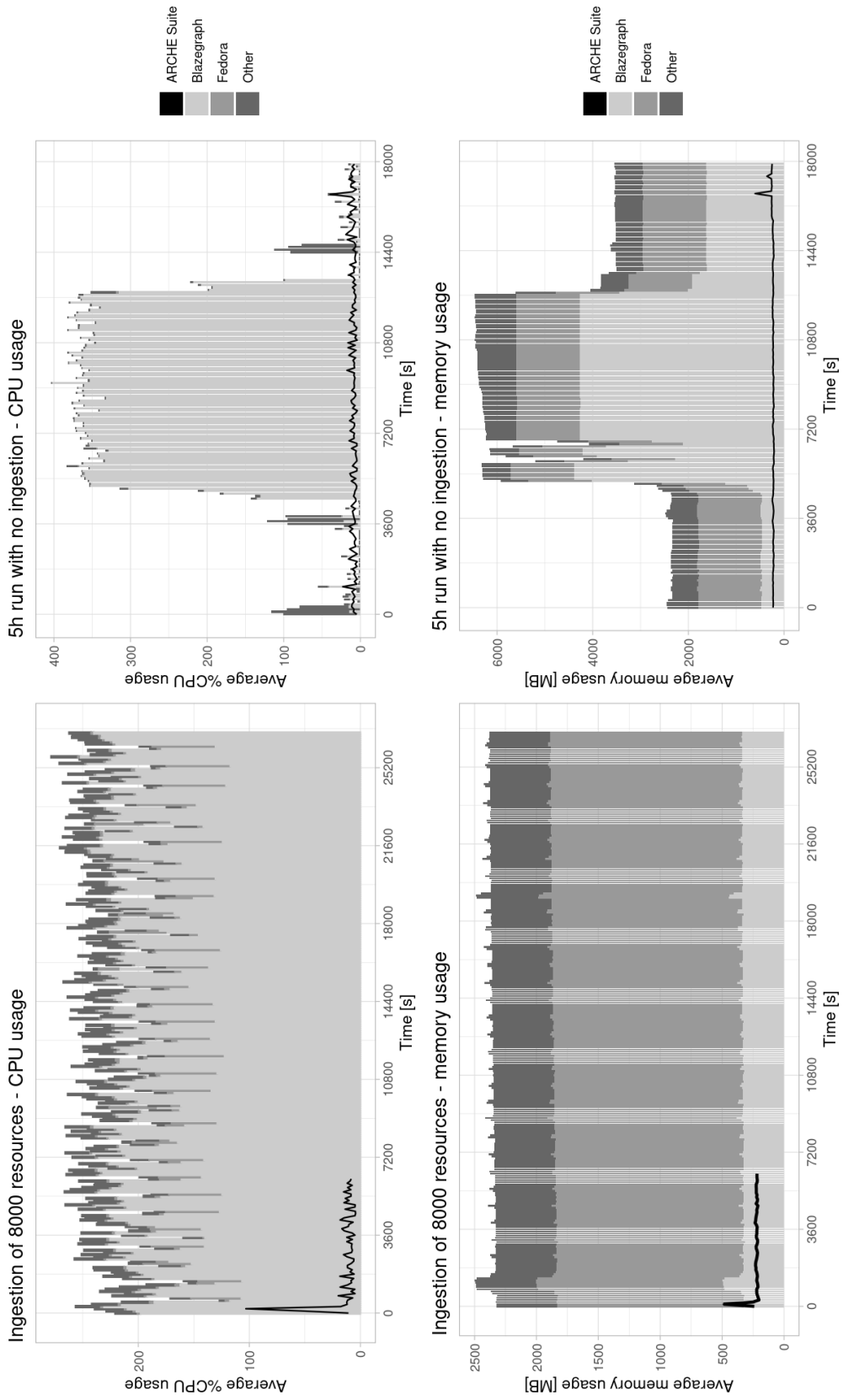


Figure 2: Comparison of hardware resources usage of the same repository implemented using Fedora Commons 4 coupled with a Blazegraph triplestore (stacked bars differentiating Fedora, Blazegraph and other components) and using the ARCHE Suite (denoted by the black line). The ingestion scenario data series for the ARCHE Suite is shorter because the ingestion finished faster.

identifier (<https://viaf/123>), is being ingested into the repository. The result is represented on the upper right-hand side of Figure 3: except for the identifiers, the RDF nodes now correspond to repository resources. Each resource was assigned an additional repository identifier (starting with <https://repoUrl/>). The identifiers of *John Doe* were imported from the RDF nodes as URIs into the repository and will be interpreted as RDF nodes upon export.

The lower part of Figure 3 represents a second ingest of another RDF graph with information about a collection with the title *Collection 2* that has the same author *John Doe*. The author *John Doe* is referenced with an already stored identifier (<https://viaf/123>) and an additional identifier is provided in the graph (<https://gnd/456>). The result from this second ingest is represented by the lower right-hand part of Figure 3: an additional repository resource for *Collection 2* was created and the additional identifier for the author is added as an URI to the already existing resource representing *John Doe*.

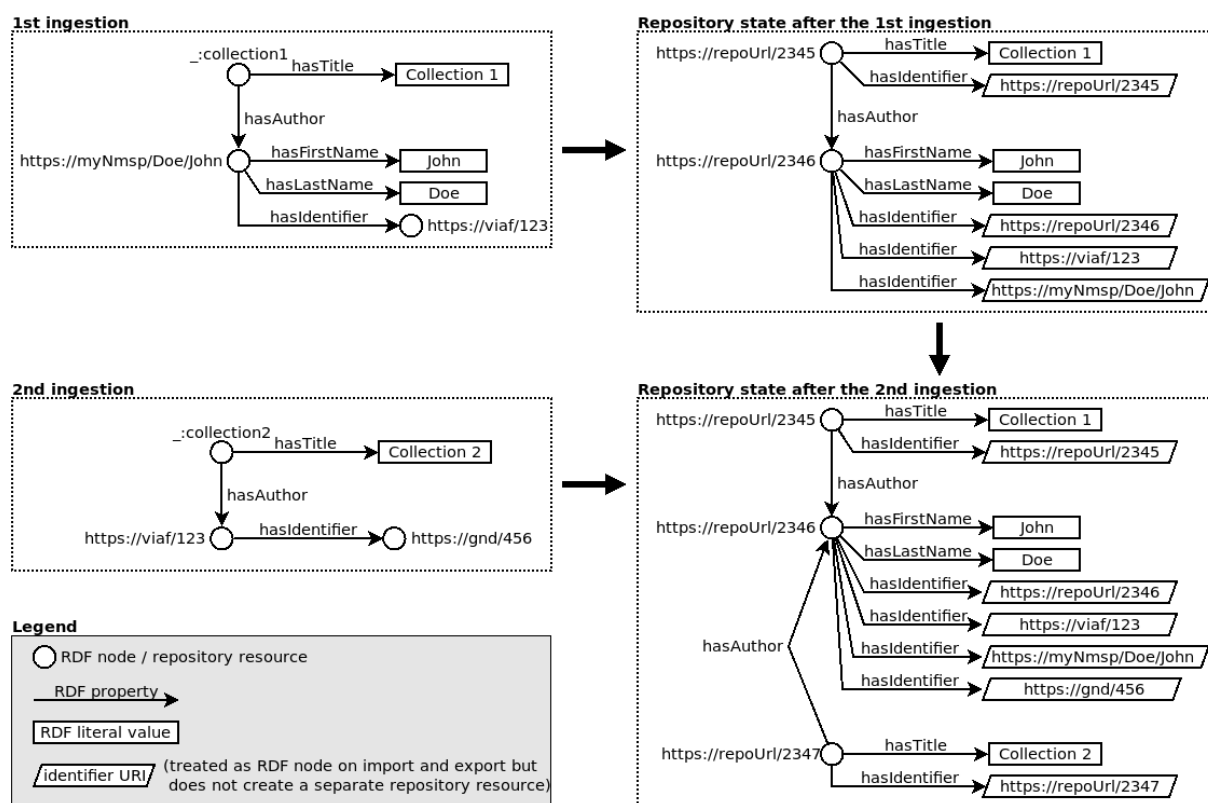


Figure 3: Example of the relation between the ingested RDF data (left) and the ARCHE Suite's internal data model (right). Identifiers are accumulated and the second ingestion does not create a new repository resource for the author but links to the already existing one.

The given example highlights that the data model used in the ARCHE Suite provides a flexible and uniform framework for handling external authoritative data. As each named entity has exactly one repository resource storing its data, e.g. information about a person (see Person *John Doe* in Figure 3), it is enough to update this resource for the change to be applied across the whole repository, i.e. all resources referring to a given person as an author (see *Collection 1* and *Collection 2* in Figure 3) do not require updating. Such an update can be done either by manual curation or by automated data retrieval from external authority files like GND⁶, VIAF⁷, ORCID⁸, GeoNames⁹, etc. We successfully employed both strategies¹⁰.

⁶ www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

⁷ viaf.org/

⁸ orcid.org/

⁹ www.geonames.org/

¹⁰ For an example of the automatic approach see github.com/acdh-oeaw/arche-ref-sources

What makes named entities handling in ARCHE Suite even more convenient is its native support for multiple identifiers per resource. The ARCHE Suite uses a dedicated and configurable RDF property to store all possible URIs, i.e. identifiers, of a given resource. In Figure 3 this property is represented as *hasIdentifier*. All identifiers stored in this RDF property are synonymous and can be used interchangeably to denote the resource. For example, if there is a repository resource with multiple identifiers like the *John Doe* resource in the lower right-hand part of Figure 3, and a new ingestion is performed denoting *John Doe* as an author, any of <https://repoUrl/2346>, <https://myNmsp/Doe/John>, <https://viaf/123> and <https://gnd/456> can be used to refer to the already existing *John Doe* repository resource in the newly ingested RDF metadata. On a conceptual level, we can say the ARCHE Suite has built-in support for the *owl:sameAs* relation, which maps all URIs being values of the above mentioned configurable RDF property to a single repository resource.

The described data model also makes the ARCHE Suite well suited to serve as an entity reconciliation back end. In fact, one of the ARCHE Suite components is a microservice providing an OpenRefine-compatible API (see left part of Figure 1)¹¹, which we are already using for curation and enrichment of metadata.

3.2 Metadata Schema and Metadata Schema Conversion

The ARCHE Suite does not enforce any particular metadata schema. The only requirement is the metadata to be expressed in RDF. The RDF predicates used for storing metadata internally managed by the repository (e.g. resource checksum, last modification date, etc.) can be adjusted in the repository configuration on run time. For example, the date of a resource's last modification can be easily set either to <http://my.own.schema#creationDate>, <http://purl.org/dc/terms/created> or even <http://fedora.info/definitions/v4/repository#created> (for direct compatibility with Fedora Commons repositories).

The OAI-PMH service shipped with the ARCHE Suite allows converting metadata into various XML-serialisable formats using a flexible templating system. We have successfully implemented conversions from our internal metadata schema to CMDI profiles as well as to the schema used by Kulturpool (the Austrian Europeana aggregator) which allows us to entirely avoid materialising metadata in specific formats (cf. OAI-DC¹², Kulturpool¹³ and CMDI profile p.1288172614023¹⁴ serialisations of the same resource).

3.3 Custom Metadata Consistency Checks

The only metadata consistency check performed automatically by ARCHE Suite is the *foreign key* constraint. As described above, all nodes of the RDF metadata graph are represented by repository resources, making it impossible to remove a repository resource that is pointed to by another resource's metadata.

All other checks have to be implemented as plugins by the repository administrator. The plugins can be written in any programming language with the AMQP message queue support¹⁵. Plugins bind to given events (before/after metadata/binary/transaction creation/modification). When an event occurs, the plugin is provided with resource metadata in the n-triples format and is expected to return metadata in the n-triples format or to raise an error. Plugins can be used both for metadata checks and enrichment as well as for synchronisation with external services (e.g. a triplestore).

The plugins system has turned out to be a very flexible and powerful tool. Dedicated plugins have been implemented for the ARCHE repository: checking metadata property cardinalities (applying different rules for resources of different RDF classes), minting PIDs, casting metadata property values to their proper RDF datatypes (including mapping string value labels to SKOS concept URIs for properties

¹¹github.com/acdh-oeaw/arche-openrefine

¹²arche.acdh.oeaw.ac.at/oaipmh/?verb=GetRecord&metadataPrefix=oai_dc&identifier=https://hdl.handle.net/21.11115/0000-000C-29F8-F

¹³arche.acdh.oeaw.ac.at/oaipmh/?verb=GetRecord&metadataPrefix=kulturpool&identifier=https://hdl.handle.net/21.11115/0000-000C-29F8-F

¹⁴arche.acdh.oeaw.ac.at/oaipmh/?verb=GetRecord&metadataPrefix=cmdi&identifier=https://hdl.handle.net/21.11115/0000-000C-29F8-F

¹⁵The are more than 20 languages with AMQP Client libraries including Java, C/C++, Python, PHP, Ruby, JavaScript/node.

with controlled vocabularies) and computing aggregated metadata property values (e.g. summary of the licence types used by resources within a collection).

3.4 Transactions Support

The ARCHE Suite provides full ACID support, although the isolation level is *read uncommitted* only. If consistency enforcement is undesired, it can be turned off by a configuration option. Importantly, all the *before event* plugins are considered part of an ACID transaction and thus, the ACID properties also extend to the plugins' actions. The transactions are backup-safe. In fact, the backup script uses its own transaction with a serialisable isolation level.

Transactions atomicity guarantees the repository can automatically get back to the pre-transaction state, i.e. perform a so-called *rollback*, if there was any error during the ingestion. This means compliance of metadata to be ingested, with the metadata schema in use can be safely checked by just performing an ingestion attempt. If there are errors, they are reported and the whole transaction is rolled back. We use such a workflow successfully for data curation and it proved to work reliably even for very large transactions, that involve an all day long ingestion of up to thousands of resources.

Due to the low isolation level, ARCHE Suite transactions have a negligible impact on the repository performance (see Figure 2) and the transaction commit is immediate. The price to be paid is a time-consuming *rollback* process taking up as much as half of the ingestion time. We did not find it troublesome in practice as the *rollback* happens only when data contains errors and the time is anyway needed to fix them.

Parallel transactions as well as parallel requests within the same transactions for faster data ingestion are also supported but discussing these complex topics in detail goes beyond the scope of this paper. More information can be found in the ARCHE Suite documentation¹⁶.

3.5 Ingestion Workflows Automation

ARCHE Suite features, especially the flexible plugins system (see Section 3.3) coupled with the ACID transactions support (see Section 3.1), allow for automated checking of input metadata compliance with the ARCHE metadata schema and performing fully automated data ingestions in a safe way. Our latest achievement is a workflow that reads metadata from TEI/XML files, maps them to the ARCHE metadata schema, and then ingests both, the source XML files and the generated metadata into the ARCHE repository¹⁷. The TEI/XML data can be stored at any place accessible via the internet, e.g. in a dedicated repository on the GitHub platform. The metadata creation and repository ingestion workflow are set up as a continuous deployment workflow using GitHub Actions¹⁸. When data is stored inside GitHub, the workflow can be automatically triggered every time a new TEI/XML data release is made. Thanks to the atomicity of the transaction described in Section 3.4, the workflow execution comes with no risk, as the transaction is rolled back whenever an error is encountered. If there is no error, the new version of the data is published without the need for any human interaction.

4 Summary

After a year and a half of using the ARCHE Suite to run the ARCHE repository (as of March 2022, over 1.9 TB of data, 132k resources, 4.5m RDF metadata triples), we can confirm it has met our expectations. It allows us to use RDF metadata as input and output format, to perform metadata enrichment and complex consistency checks within the repository software, as well as to avoid duplicating metadata by materialising various metadata formats. Notably, using the ARCHE Suite has significantly reduced server resources consumption compared to the previous solution based on Fedora Commons 4 coupled with a Blazegraph triplestore. We are determined to develop the ARCHE Suite further and seek for cooperation with other CLARIN partners.

¹⁶acdh-oeaw.github.io/arche-docs/aux/parallel_ingestion.html

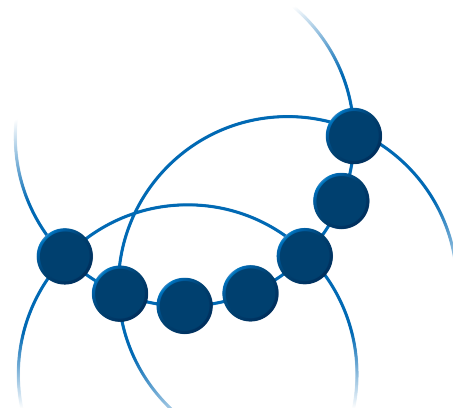
¹⁷For a practical use case see e.g. github.com/acdh-oeaw/kraus-static/actions

¹⁸docs.github.com/en/actions/learn-github-actions

References

- Tim Berners-Lee. 2009. Linked data.
- Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. Cmd: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.
- CERN, Northwestern University, and contributors. 2021. Inveniordm - reference documentation: Metadata reference.
- DCMI Usage Board. 2020. DCMI metadata terms.
- DuraSpace. 2021. Dspace 7.x documentation - linked (open) data.
- Theo Haerder and Andreas Reuter. 1983. Principles of transaction-oriented database recovery. *ACM Computing Surveys*, 15(4):287–317.
- Timothy Holborn. 2014. What is 5 star linked data?
- Lars Holm Nielsen. 2019. Inveniordm: a turn-key open source research data management platform.
- Institute for Quantitative Social Science. 2021. Dataverse user guide - supported metadata export formats.
- Gary King. 2007. An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods and Research*, 36:173–199.
- Mikael Nilsson, Andy Powell, Pete Johnston, and Ambjörn Naeve. 2008. Expressing dublin core metadata using the resource description framework (RDF).
- Jenn Riley. 2010. Seeing standards: A visualization of the metadata universe.
- MacKenzie Smith, Mary Barton, Mick Bass, Margret Branschofsky, Greg McClellan, Dave Stuve, Robert Tansley, and Julie Harford Walker. 2013. Dspace. an open source dynamic digital repository. *D-Lib Magazine*, 9(1).
- The Fedora Leadership Group. 2016. Fedora and digital preservation.
- Martina Trognitz and Matej Ďurčo. 2018. One schema to rule them all. the inner workings of the digital archive ARCHE. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 71(1):217–231, July.
- W3C, Steve Harris, Andy Seaborne, and Eric Prud’hommeaux. 2013. Sparql 1.1 query language.
- W3C, Richard Cyganiak, David Wood, Markus Lanthaler, Graham Klyne, Jeremy J. Carroll, and Brian McBride. 2014. Rdf 1.1 concepts and abstract syntax.
- 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(1), March.

CLARIN



Common Language Resources and Technology Infrastructure

Linköping Electronic Conference Proceedings
eISSN 1650-3740 (Online) • ISSN 1650-3686 (Print)
ISBN 978-91-7929-444-1

189
2022

Front Cover Illustration:
Picture Composition by CLARIN ERIC