CLARIN Annual Conference Proceedings

2023

Edited by

Krister Lindén, Jyrki Niemi, and Thalassia Kontino

16 – 18 October 2023
Leuven, Belgium

Please cite as:
CLARIN Annual Conference Proceedings, 2023. ISSN 2773-2177 (online).
Eds. Krister Lindén, Jyrki Niemi, and Thalassia Kontino.
Leuven, Belgium, 2023.

Programme Committee

Chair:

- Krister Lindén, University of Helsinki (FI)

Members:

- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies (IS)
- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Cristina Grisot, University of Zurich (CH)
- Eva Hajičová, Charles University Prague (CZ)
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Gijsbert Rutten, Leiden University (NL)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič, University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven (BE)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Joshua Wilbur, Univerity of Tartu (EE)
- Andreas Witt, University of Mannheim (DE)
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University (ZA)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Starkaður Barkarson, IS
- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Cristina Grisot, CH
- Eva Hajičová, CZ
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Maciej Piasecki, PL
- Stelios Piperidis, GR
- Gijsbert Rutten, NL
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Vincent Vandeghinste, BE
- Tamás Váradi, HU
- Joshua Wilbur, EE
- Andreas Witt, DE
- Friedel Wolff, ZA
- Martin Wynne, UK

Subreviewers:

- Ilze Auzina, LV
- Federico Boschetti, IT
- Riccardo Del Gratta, IT
- Amelie Dorn, AT
- Maria Gavriilidou, GR
- Luís Gomes, PT
- Marissa Griesel, ZA
- Kinga Jelencsik-Mátyus, HU
- Mateja Jemec Tomazin, SI
- Fahad Khan, IT
- Penny Labropoulou, GR
- László János Laki, HU
- Kristine Levane-Petrova, LV
- Noémi Ligeti-Nagy, HU
- Amália Mendes, PT
- Hannes Pirker, AT
- Valeria Quochi, IT
- João Rodrigues, PT
- Rodrigo Santos, PT
- João Silva, PT
- Juan Steyn, ZA
- Benito Trollip, ZA
- Menno van Zaanen, ZA

CLARIN 2023 submissions, review process and acceptance

- Call for abstracts: 23 January 2023 first call published on CLARIN website, disseminated, and submission system open
- Submission deadline: 28 April 2023
- In total 52 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 9 June 2023
- Notifications to authors: 30 June 2023
- 37 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2023/clarin-annual-conference-2023.

Table of Contents

Meta Data and Annotation

ParlaMint

Tools

Posters

A Spoken Academic Belgian Dutch Corpus

Vincent Vandeghinste
Instituut voor de Nederlandse Taal
vincent.vandeghinste@ivdnt.org

Jolien Mathysen
KU Leuven, Belgium
jolien.mathysen@kuleuven.be

Elke Peters
KU Leuven, Belgium
elke.peters@kuleuven.be

Patrick Wambacq
KU Leuven, Belgium
patrick.wambacq@kuleuven.be

Abstract

We present the Spoken Academic Belgian Dutch (SABeD) corpus. It was compiled from selected first bachelor academic lectures in higher education institutions in Flanders, as students indicate that the language used in such lectures is one of the hurdles for comprehension and academic success. We first applied speech recognition on these lectures, and then applied manual utterance segmentation, and manual correction of the automated transcription. The resulting text is processed with the FROG language analyser and will be made searchable through a CLARIN website as soon as all manual editing is done.

1 Introduction

In higher education, students are confronted with academic language use, with which they are often not familiar. Since academic language skills are a necessary condition for study success, higher education institutions in Flanders and the Netherlands focus on language support for students. In many institutions, these efforts evolved into formal, embedded language policies, but research into their implementation is limited (Bonne & Casteleyn, 2022).

The number of international students pursuing higher education in Flanders is estimated around 2500 per year (Deygers & Malone, 2019). Research (Deygers, 2017; Deygers et al., 2017) shows that Dutch language learners struggle with academic spoken Dutch, even when they passed the university entrance language tests, ITNA[1] or CNaVT[2]. Although academic listening is part of the test, learners indicate that the listening tasks in the language entrance tests are easier than actual lectures (Deygers et al., 2018). Linguistic features of the listening task in the test have not been empirically validated because of the lack of a corpus of spoken academic Dutch. This is one of the main reasons for building a corpus of spoken academic Belgian Dutch, which consists of (recordings of) academic lectures. Lectures are typical of higher education, and due to the covid pandemic, recorded video lectures are available in abundance. The corpus contains a mix of written-to-be-read and spontaneous spoken language.[3] This paper presents how such a corpus was created.

[1] https://www.itna.be/

[2] https://cnavt.org/

[3] In analogy with CGN,(Oostdijk et al., 2002) we consider our corpus as a *spoken corpus*. The initial collection of 1028 recordings did include a number of pre-recorded lectures and knowledge clips, in which the lecturers prepared and read out their text. However, in the selection of the 200 recordings for the final corpus, live lectures taught on campus were given priority. Additionally, other spoken corpora, such as the CGN also feature a variety of (semi-)structured instances of speech (e.g. interview, news bulletins, masses, formal speeches and even recited texts) (cf. https://ivdnt.org/images/stories/producten/documentatie/cgn_website/doc_Dutch/topics/overview.htm#inleiding).

2 Related Work

Three of the most notable corpora featuring spoken academic language currently in existence are the T2K-SWAL (*TOEFL 2000 Spoken and Written Academic Language*) corpus (Biber et al., 2002), BASE (*British Academic Spoken English*) corpus (Thompson & Nesi, 2001), 2001) and the CGN (*Spoken Dutch Corpus* Oostdijk et al., 2002. The spoken component of the T2K-SWAL corpus includes 1.7 million words recorded at four different American universities. The largest part of this (1.2 million words) were taken from 176 class sessions, while the remaining 50,000 came from office hours (Biber et al., 2002). BASE consists of 160 lectures and 40 seminars recorded at the University of Warwick and the University of Reading between 2000 and 2005. This 1,186,290 token corpus was compiled from four disciplinary sub-corpora: Arts and Humanities, Life and Medical Sciences, Physical Sciences, and Social Sciences. Except for Physical Sciences, each sub-corpus contains 40 lectures and 10 seminars. The BASE corpus is the most recent corpus of academic spoken English (Thompson & Nesi, 2001). CGN is a balanced corpus with several variants of spoken Dutch (from read-aloud text to spontaneous conversations, from Belgium and the Netherlands), and which contains 30,917 words from university lectures (Oostdijk et al., 2002). However, a domain specific spoken Dutch corpus like the one we propose in this paper was until recently not yet publicly available.

3 Corpus Compilation

In the corpus compilation stage, we selected academic lectures, because these constitute the predominant form of instruction in higher education institutions in Flanders, especially in the first bachelor year. Lectures are defined as instructional discourse given before an audience of at least 40 students, in which the lecturer is the dominant speaker and the level of interactivity is modest to low. We chose lectures for first year bachelor students, as both native speakers and foreign learners of Dutch indicate the language used in lectures as one of the hurdles for comprehension and academic success (Deygers, 2017; Deygers et al., 2017). First year bachelor lectures also constitute the first encounter of the target group (i.e., Flemish first bachelor students and international students commencing university education in Belgian-Dutch) of our corpus with spoken academic Dutch. As such, these lectures make up a solid base for our corpus compilation, especially considering that we cannot be certain if and to what degree the language of lectures in later bachelor and master years differs from that in the first bachelor. It is also important to take into account the primary pedagogical goal of the corpus, i.e., developing learning materials for students entering Flemish higher education.

To ensure that the corpus is both representative and has sufficient power to make statistical inferences, lectures from a considerable number of lecturers need to be included (Biber, 1993).

After the example of the BASE corpus, the selection of lectures is further informed by academic division (biological and health sciences, humanities and arts, physical sciences and engineering, social sciences and education). Within each academic division, a broad range of disciplines is covered. We aim to compose a corpus that is sufficiently representative for the purpose of composing a word list of spoken academic Dutch. At the same time, the corpus should contain sufficient data to obtain standards for more specific academic divisions.

The selection of lectures for the transcription stage was impeded by the fact that, due to technical reasons, it was not possible to automatically download lectures from the video platforms used by Flemish universities. All lecturers/professors had to be contacted individually and all lectures had to be downloaded and processed one-by-one before they could be added to the corpus. Informed consent was obtained and metadata was collected (e.g. speaker data such as age, gender, teaching experience, place of birth). We obtained 1028 lectures, which is more lectures than the existing corpora of spoken academic English such as the T2K-SWAL Corpus (176 lectures; (Biber et al., 2002)) or BASE (160 lectures; Thompson and Nesi, 2001.

Finally, we also collected written course materials, to increase the accuracy and performance of the speech recognition system (cf. section 4.1).

4 Transcription

We only manually transcribed part of the lectures, instead of transcribing a considerably smaller number of entire lectures. Transcribing only the first 25 and last 5 minutes of the lectures has the practical benefit that any differences between institutes and disciplines in length of lectures are eliminated, ensuring a well-balanced corpus.

4.1 Automated speech recognition

Speech recognition was performed with an ASR system tuned for Belgian Dutch (Van Dyck et al., 2021), which is Kaldi-based (Povey et al., 2011). The Kaldi toolkit makes use of state-of-the-art deep neural networks. The new acoustic model was trained on data from the Spoken Dutch Corpus (Oostdijk et al., 2002) and tested using the N-best benchmark (Kessens & van Leeuwen, 2007). The output of the ASR system consists of `ctm` files,[4] which contain time stamps for each recognized word and a confidence level for each word.

An alternative for this system would be to use the recently emerging end-to-end systems, but these do not allow independent training of language models, which is exactly what we want in this project. The additional text material from textbooks and course materials will be used to improve the lexicon and language model independently from the acoustic model. This latter effort is still ongoing.

The *raw ASR* output is inserted into an ELAN file as a separate tier. ELAN (Wittenburg et al., 2006) is well known software for linguistic annotation of audio-visual material.

4.2 Manual Post-editing

A first step in manual post-editing consists of applying **utterance segmentation**.[5] This segmentation annotation is registered in a separate ELAN tier. It entailed the placement of boundaries which indicate the start and ending of a piece of transcription in the audio signal. The unit which demarcates these boundaries is the chunk, which is defined as a speech fragment which lasts about 2-3 seconds and which is delimited on both sides by a (short) audible and visible pause. Chunks can, but do not need to, correspond to sentences or phrases. Chunks longer than 3 seconds are allowed (e.g. multiple co- or subordinated clauses, long enumerations), however, if chunks last longer than 6 seconds, they are split up in front of a conjunction or where a comma would appear in written language. Any student interactions or substantial background noises (e.g. an opening door) were also isolated in segments and tagged in a separate ELAN tier.

The *raw ASR* tier is combined with the *manual segmentation* tier into a *segmented ASR tier*.

The second step consists of **correcting the automated transcription** at the segment level. Annotators manually correct the *segmented ASR* tier using a transcription protocol that was based on that of the Spoken Dutch Corpus (Corpus Gesproken Nederlands, Oostdijk et al. 2002) and put the correction into the *manual transcription* tier. More specifically, there are two phases involved in the manual correction of the transcriptions. During the first phase, transcriptions are corrected at the orthographic level. This entails that the spelling is standardised and punctuation is added. Interjections and words from languages other than Dutch are annotated using designated codes. Students speech is cut out because it is too difficult to track down students and get their consent and personal data (e.g., names of lecturers, students, courses) are anonymised because of GDPR. During the second phase, transcriptions are checked at the acoustic level. This includes the annotation of reductions, dialect words, slips of the tongue, aborted words and sentences, unintelligible pieces of speech, and noises made by the speakers (e.g., coughing or sneezing). Manual correction is significant in that it allows calculating Word Error Rate of the ASR by comparing the two tiers.

[4] `ctm` stands for time-marked conversation file.

[5] This is done to make post-editing easier and faster. If we would not have applied segmentation, ASR correction would necessarily have to be performed at the word level, which would be cumbersome as ASR errors can span over word boundaries, requiring annotators to not only correct the transcript but also manually manipulate the time stamps.

5 Further processing

A first batch of manually corrected transcriptions has been fed back into the ASR, with a second batch of videos being well on its way to being manually corrected. The corrected first batch combined with the use of written text material, should improve the accuracy of the ASR considerably.

This first batch has been post-processed using the CLARIN tool FROG (Van den Bosch et al., 2007), which results in the following analyses: tokenization, part-of-speech tagging, lemmatization, morphological segmentation, dependency parsing and named entity labeling. FROG outputs FoLiA format (van Gompel & Reynaert, 2013), an xml format made for linguistic annotations.

These FoLiA data have been uploaded to the CLARIN Autosearch engine (de Does et al., 2017) for annotated corpora, making this part of the corpus now searchable with Corpus Query Language (CQL) and sharable with other CLARIN users.

6 Conclusions and future work

We have presented the corpus collection efforts for a corpus of spoken academic Belgian Dutch. Once all data is processed, combining the metadata with the linguistic annotations in TEI format will allow more fine-grained querying of the corpus, not only on linguistic criteria but also on metadata criteria, and the corpus will be made available in a Blacklab (de Does et al., 2017) corpus query engine for all CLARIN users.

While ASR speeds up the manual transcription, it is clear that a general domain ASR system does not contain a specialized vocabulary like is being used in academic lectures, and therefore tuning the vocabulary and language model of the ASR system towards the specific domains is expected to greatly improve ASR accuracy and reduce post-editing effort, which should result in a speedier post-editing process.

Validation of word lists and language tests for academic Belgian Dutch was one of the main reasons for collecting the corpus, but is still future work. The development of an academic spoken word list will be based on the frequency and range of the words in the corpus (Dang et al., 2017; Szudarski, 2017) with the lemma as counting unit. This functionality is included in the Blacklab environment. To determine which words can be considered academic words, the frequency list will be compared to the word list of Tiberius and Schoonheim, 2013. Words not occurring in that list are potential candidates for the spoken academic vocabulary list, depending on their frequency and distribution in the corpus. We will distinguish proper names, general academic words and domain-specific words. As in the English Academic Spoken Word list, (Dang et al., 2017) we will divide the list into sublists of 50 words, based on their frequency.

We will also develop a frequency-based spoken vocabulary test targeting students' aural recognition of academic words. The test will be divided into test sections, corresponding to the sublists of the frequency list. The test will have an online multiple choice format and students will be provided with the spoken form of the word and will have to tick off the correct option. The first test version will be piloted with a small group of Dutch-speaking students (n=30) before the start of the actual larger scale validation process.

Of course, once the corpus has been made available to researchers, a multitude of other uses and applications can be envisaged, such as comparisons at lexical, syntactic and other levels with other (spoken and/or written) Dutch corpora.

Acknowledgments

The SABeD project is funded by KU Leuven Internal Funding, Research Project 3H200610, with additional support from the Instituut voor de Nederlandse Taal.

References

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, *8*, 243–257.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, *36*, 9–48.

Bonne, P., & Casteleyn, J. (2022). Taalbeleid en taalondersteuning: Op zoek naar een gedeelde basis en strategie voor implementatie. *Tijdschrift voor Onderwijsrecht en Onderwijsbeleid*, *4*, 279–293.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, *67(4)*, 959–997.

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating Research Environments with BlackLab. In *CLARIN in the Low Countries*. Ubiquity Press.

Deygers, B. (2017). Validating university entrance policy assumptions. Some inconvenient facts. In E. Gutiérrez Eugenio (Ed.), *Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference* (pp. 46–50). Cambridge: ALTE.

Deygers, B., & Malone, M. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, *36(3)*, 347–368.

Deygers, B., Van den Branden, K., & Peters, E. (2017). Checking assumed proficiency: comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, *32*, 43–56.

Deygers, B., Van den Branden, K., & Van Gorp, K. (2018). University entrance language tests: A matter of justice. *Language Testing*, *35*, 449–476.

Kessens, J. M., & van Leeuwen, D. A. (2007). N-best: the northern- and southern-Dutch benchmark evaluation of speech recognition technology. *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, 1354–1357. http://www.isca-speech.org/archive/interspeech%5C_2007/i07%5C_1354.html

Oostdijk, N., Goedertier, W., van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & Baayen, H. (2002). Experiences from the spoken Dutch corpus project. *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Szudarski, P. (2017). *Corpus linguistics for vocabulary. A guide for research*. Routledge.

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, *5*, 263–264.

Tiberius, C., & Schoonheim, T. (2013). *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge.

Van den Bosch, A., Busser, G., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114).

Van Dyck, B., BabaAli, B., & Van Compernolle, D. (2021). A Hybrid ASR System for Southern Dutch. *Computational Linguistics in the Netherlands Journal*, *11*, 27–34.

van Gompel, M., & Reynaert, M. (2013). FoLiA: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, *3*, 63–81.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

NGT-HoReCo and GoSt-ParC-Sign: Two new Sign Language - Spoken Language parallel corpora

Mirella De Sisto, Dimitar Shterionov
Tilburg University, the Netherlands
M.DeSisto/D.Shterionov
@tilburguniversity.edu

Lien Soetemans
KULeuven, Belgium
lien.soetemans@kuleuven.be

Vincent Vandeghinste
Dutch Language Institute
Leiden, the Netherlands
vincent.vandeghinste@ivdnt.org

Caro Brosens
Vlaams Gebarentaalcentrum
Antwerp, Belgium
caro.brosens@vgtc.be

Abstract

Language technology targeting both signed and spoken languages is extremely limited. This is partly due to the scarce availability of good quality signed language data and of signed and spoken parallel corpora. In this paper we introduce two projects which aim at reducing the gap between spoken-only language technology and more inclusive language technology for both signed and spoken languages by creating two parallel corpora with a sign language on one side and a spoken language on the other: the Dutch - Sign Language of the Netherlands Hotel Review Corpus (NGT-HoReCo) and the Gold Standard Parallel Corpus of Signed and Spoken Language (GoSt-ParC-Sign). Both corpora are or will be made available through the CLARIN infrastructure.

1 Introduction

In Europe about half a million people have a Sign Language (SL) as their main or preferred means of communication (Pasikowska-Schnass, 2018). Nevertheless, when talking about language technology, SL technology is extremely lagging behind in comparison to the tools available for spoken languages (Vandeghinste et al., 2023). One of the reasons is the scarcity of data (for a detailed overview of data-related challenges, see De Sisto et al., 2022; Vandeghinste et al., forthcoming); this is partially due to the fact that SLs do not have a widely-used written form used by Deaf communities, hence spontaneous written data is not an option (as it is the case for many spoken languages). Data collection and data storage also face a number of challenges, such as GDPR restrictions, difficulties in recruiting participants, etc.

The majority of SL data comes in the form of videos. To date there is no automatic tool able to annotate or translate SL videos (Morgan et al., 2022; Vandeghinste et al., 2023), which means that any of these processes relies on very time-consuming manual work; consequently, the amount of annotations or translations available is scarce.

In addition to that, often the quality of the data available is rather problematic (Vandeghinste et al., forthcoming). Most of the ML-readable SL datasets are news broadcast original spoken language interpreted by hearing interpreters, which is rather problematic in terms of the quality of the data: firstly, in those cases SL is the target language of interpreting which often occurs simultaneously, hence, is both influenced by the source language as well as affected by the interpreting process; secondly, most hearing interpreters do not use a SL as their main or preferred means of communication (the exception being interpreters being CODA's – Children of Deaf Adults – and some other specific cases); consequently, they can be considered L2 signers.

In this paper we present two recent projects which address the lack of good quality data by providing two parallel corpora of signed and spoken language data: the Dutch - Sign Language of the Netherlands Hotel Review Corpus (NGT-HoReCo) which consists of a parallel dataset of hotel reviews in written English, written Dutch (translations of the original English by a professional translation service in Dutch), and Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT) videos; the Gold Standard Parallel Corpus of Signed and Spoken Language (GoSt-ParC-Sign), a golden standard dataset of

semi-spontaneous Flemish Sign Language (Vlaamse Gebarentaal) (VGT) videos translated into written Dutch. Such datasets, that include SL data produced by native signers and have been collected in a way that suits their use in ML applications, have the potential to stimulate the advancements in the field of SL technology through both high-quality data for training models as well as a gold standard for testing.

2 NGT-HoReCo

The NGT-HoReCo project took place between January and March 2023. The corpus consists of hotel reviews in written English, translated into written Dutch and into NGT videos. The Dutch text was produced by translating Booking.com hotel reviews from English to Dutch. These reviews are publicly available on Kaggle. [1] The English-Dutch translations were produced by a professional translation company which used automatic translation (generated by DeepL) following an in-depth human post-editing. Dutch to NGT translations were performed by six deaf professional translators. Relying on deaf and not hearing interpreters we ensured that (i) there is as little as possible interference of the source language (Dutch) and (ii) the signing is authentic, i.e. produced by a native (L1) signer. The corpus consists in 283 reviews: 19,950 words in the English source, 21,825 words, on the Dutch text side, and 213.18 minutes on the NGT video side. The advantage of providing data focusing on a single domain, i.e. hospitality, allows to have recurrent topics and signs in different possible combinations and to account, to a certain extent, for inter and intra signer variation.

Figure 1 shows an example of the parallel texts and video. One folder contains all videos. An excel file contains the original English text, a Dutch translation obtained with machine translation, the Dutch translation produced by the translation company; the last two columns contain the video identifier and the signer identifier, respectively.

The hotel was beautiful and the staff was awesome. One of the best beaches in mexico	Het hotel was prachtig en het personeel was geweldig. Een van de beste stranden in Mexico	Het hotel was prachtig en het personeel was geweldig. Een van de beste stranden in Mexico.	NGT-HoReCo_89	P3

Figure 1: Example from NGT-HoReCo

The corpus is available at http://hdl.handle.net/10032/tm-a2-w2 under CC BY-NC license. A CMDI record has been made which should be harvested by the CLARIN Virtual Language Observatory to ensure findability of the corpus. The corpus is also available through the European Language Grid at https://live.european-language-grid.eu/catalogue/corpus/21535.

[1] https://www.kaggle.com/datasets/datafiniti/hotel-reviews

3 GoSt-ParC-Sign

The GoSt-ParC-Sign project started in February 2023 and will be ongoing until January 2024. The corpus will contain videos of spontaneous and semi-spontaneous VGT produced by deaf individuals who use VGT as their main or preferred means of communication for deaf or signing audience. The project consists of three phases.

In the first phase, we identified roughly ten hours of publicly available (semi-)spontaneous VGT videos. These videos cover different topics and genres, such as five hours of free conversation, one and a half hour of panel discussion about linguistic change in the community, over two hours of a deaf-lead talk, a game show to celebrate 15 years of recognition for VGT, and 45 minutes of semi-spontaneous vlogs about typical language uses in VGT. Currently, informed consents for the public availability of the videos are being collected from the video owners. In addition, we recruited a mixed team of deaf and hearing professional VGT translators; having both deaf and hearing translators makes sure that the content of the source is preserved, and ensures good quality of the target translation.

The second phase will focus on translating the VGT videos into written Dutch text. Translations will be organised in ELAN (Sloetjes & Wittenburg, 2008), which allows multiple annotation tiers synchronised with the video timeline. A 'Translation' tier will contain the written Dutch translation in each ELAN Annotation Format (EAF) file of each video (an example of the format is provided in Figure 2). Having files in EAF can serve for linguistic research; in addition, this format can be easily adjusted into an ML-suited format with the framework proposed in De Sisto et al. (2022).

In the third phase, the coordinators of the translation team together with members of the VGT Deaf community will perform quality control of the translations produced.

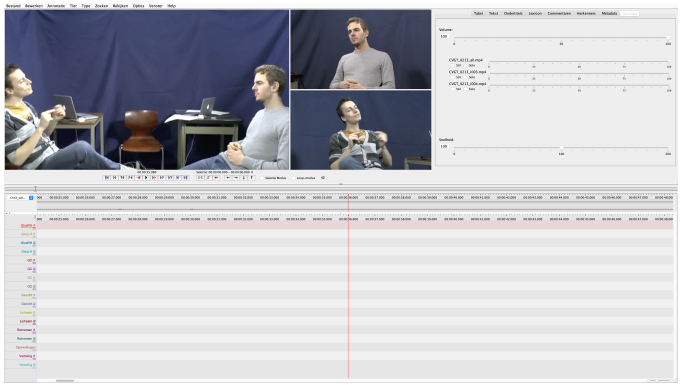


Figure 2: Example of GoSt-ParC-Sign's data format

The corpus will be made publicly available, under CC BY license, at the Instituut voor de Nederlandse Taal (INT), which ensures long-term availability. The metadata will also be published in CMDI formats for harvesting by the CLARIN infrastructure.

4 Use-case

Various SL datasets have been collected over the years, e.g. CorpusNGT (Crasborn et al., 2008) or DGSKorpus (Prillwitz et al., 2008). However, such datasets are not particularly suited for machine learning or deep learning applications, and require substantial processing prior to building language technology for SLs (De Sisto et al., 2022; Vandeghinste et al., forthcoming). Within these two projects we take this into account. Along with the open distribution of these data sets (making them available for the wider research community), the quality of the data (professional translations, involvement of native signers for translation and validation, etc.), and the different (identifiable) domains, they have been collected in a way that suits their use in ML applications, and thus have the potential to stimulate the advancements in the field of SL technology through both high-quality data for training models as well as a gold standard

for testing. For example, we have already initiated the further development of the NGT-HoReCo corpus to cover VGT, different type of annotations, pose estimates, etc., to facilitate ML and DL applications. Within the GoSt-ParC-Sign we will use ELAN, following standards to allow for the straightforward use of the data by linguists (familiar with ELAN and the EAF) as well as DL/ML practitioners using tools such as De Sisto et al., 2022.

5 Conclusion

In this paper we have introduced two SL data collection projects which aim at supporting advances in more inclusive language technology which also targets SLs. The very recently concluded NGT-HoReCo project led to the creation of a Dutch - NGT parallel corpus which contains 283 hotel reviews in written English, Dutch and NGT videos. The GoSt-ParC-Sign project is still ongoing and aims at creating a parallel corpus of authentic VGT videos and a translation into written Dutch. The creation of similar parallalel data is fundamental for supporting research and developments into fields such as SL translation, recognition and processing.

Acknwoledgements

The NGT-HoReCo project has been funded by the SRIA Contribution Projects of the ELE 2 project. The GoSt-ParC-Sign project has been awarded the EAMT Sponsorship of Activities 2022 and is partially funded by the SignON project, funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101017255.

References

Crasborn, O., Zwitserlood, I., & Ros, J. (2008). Het Corpus NGT. Een digitaal open access corpus van filmpjes en annotaties van de Nederlandse Gebarentaal. Nijmegen: Centre for Language Studies, Radboud University. https://www.corpusngt.nl/

De Sisto, M., Vandeghinste, V., Egea Gómez, S., De Coster, M., Shterionov, D., & Saggion, H. (2022). Challenges with sign language datasets for sign language recognition and translation. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2478–2487. https://aclanthology.org/2022.lrec-1.264

Morgan, H. E., Crasborn, O., Kopf, M., Schulder, M., & Hanke, T. (2022). Facilitating the spread of new sign language technologies across Europe. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, J. Mesch, & M. Schulder (Eds.), *Proceedings of the LREC2022 10th workshop on the representation and processing of sign languages: Multilingual sign language resources* (pp. 144–147). European Language Resources Association (ELRA). https://www.sign-lang.uni-hamburg.de/lrec/pub/22026.pdf

Pasikowska-Schnass, M. (2018). *Sign languages in the EU* (tech. rep.). European Parliamentary Research Service. http://www.europarl.europa.eu/RegData/etudes/ATAG/2018/625196/EPRS_ATA(2018)625196_EN.pdf

Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., & Schwarz, A. (2008). DGS corpus project–development of a corpus based electronic dictionary German Sign Language/German. *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 159.

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category: ELAN and ISO DCR. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/208_paper.pdf

Vandeghinste, V., De Sisto, M., Kopf, M., Schulder, M., Brosens, C., Soetemans, L., Omardeen, R., Picron, F., Van Landuyt, D., Murtagh, I., Avramidis, E., & De Coster, M. (2023). *Report on Europe's Sign Languages* (tech. rep.). European Language Equality D1.40.

Vandeghinste, V., Sisto, M. D., Gómez, S. E., & Coster, M. D. (forthcoming). *Challenges with sign language datasets*.

Teaching Syntax with Clarin Corpora and Resources

Antonio Balvet
UMR 8163 – STL "Savoirs Textes Langage" – F-59000
Lille University, France
antonio.balvet@univ-lille.fr

Abstract

The recent COVID-19 pandemic has brought online learning to the forefront for learners and teachers. As a consequence, the demand for self-paced and adaptive learning resources has reached unprecedented levels. Prior to the virus outbreak and consecutive lockdowns, universities had been using Moodle (and other SCORM[1] compliant platforms) as a Learning Management System (LMS), which has helped make the transition from on-site to online learning. But teachers still have had to face the challenge of designing and implementing assessment activities in the form of self-correcting activities (true/false, multiple answer questions, mark the words, fill in the blanks questions, etc.), instead of plain printed quizzes and tests. This step has proved to be a major hurdle since designing, and most of all, manually editing formative and evaluative assessment activities is a very labor-intensive task. In this article, we present a framework that builds upon corpora and resources available from the LINDAT / CLARIAH-CZ Data & Tools platform in order to generate quizzes and other activities related to syntax, for the Moodle platform. After some background on using Natural Language Processing (NLP) and electronic corpora for teaching syntax, we present our corpus-to-quiz processing chain, and we outline preliminary results on deploying automatically generated French syntax quizzes in the classroom.

1 Introduction

The recent COVID-19 pandemic has emphasized the necessity of self-paced and adaptive learning resources. Even though universities around the world had been using e-learning platforms prior to this event, teachers were still confronted with a very labor-intensive task, since designing and editing self-correcting assessment activities for potentially large groups of learners, in a distance-learning context, proved very time-consuming. Moreover, designing and implementing such assessment activities by hand is both error-prone and subjective, by nature.

In this article, we present a solution to optimize manual labor by relying on publically-available corpora. In the first section, we outline projects that have been using NLP solutions for teaching syntax. In the second section, we present our corpus-to-quiz processing chain, which ingests annotations present in corpora available from the LINDAT / CLARIAH-CZ Data & Tools platform, to generate syntax quizzes. Lastly, we report preliminary results on deploying such automatically generated quizzes, both for distance and on-site learning. Our presentation is centered on French, although the principle presented here is applicable to any Universal Dependencies CONLL-U formatted corpus, with minor adjustments.

1.1 Background: using parsers and annotated corpora for linguistic exercises and activities

Our corpus-to-quiz processing chain aims both at reducing manual edition to a minimum, and at overcoming the subjectivity (and errors) associated with manually-created exercises. Other projects have tried to address exactly those issues, in the past, such as (Bick, 2001, 2004; Uibo & Bick, 2005; Wijlff, 2006),

[1] Sharable Content Object Reference Model.

in the framework of the VISL Corpus project[2]. This project was based on a Categorial Grammar parser architecture, tailored to different languages. Based on this CG parser, large syntactically parsed corpora were set up, which allowed VISL consortium members to implement a platform very similar to the well known "Sketch Engine" (Kilgarriff et al., 2008). In addition to syntax-aware concordancers, members of the VISL consortium also devised an array of gamified exercises, based on the CG-parsed corpora, for different languages.[3]

More recently, other projects have integrated high-precision and robust parsers to automatically generate grammar exercises for French: (Colin, 2020; Perez-Beltrachini et al., 2012), in the framework of the LORIA-led METAL project.[4]

Our approach distinguishes itself from the aforementioned projects in that it builds upon manually-verified syntactic annotations, taken from reference corpora –such as the French Treebank or Sequoia– in order to generate quiz questions, which are ready to integrate into LMS[5] platforms such as Moodle[6]. Moreover, our approach targets undergraduate students, while the METAL project, for example, targets primary school pupils. Therefore, our approach focuses on the exercise generation aspect, for an audience of young adults; all authentication procedures and learning analytics logging are handled by the particular LMS being used.

2 A corpus to quiz processing chain

Our processing chain for generating self-correcting quizzes on French syntax relies on CONLL-U formatted corpora. At the time of writing, the French Treebank (FTB) (Abeillé et al., 2003) and the Sequoia corpus (Candito et al., 2014) are the only reference corpora, annotated following the Universal Dependencies guidelines (De Marneffe et al., 2021), available for French.[7] Our corpus-to-quiz processing chain is outlined in figure 1.

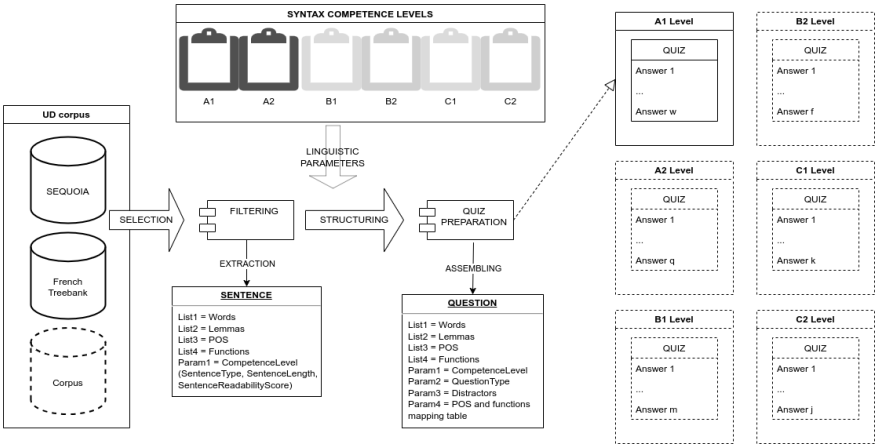


Figure 1: The corpus to quiz processing chain

Our tool is primarily targeted at university students attending introductory courses on syntax, as part of a curriculum in linguistics. We therefore need a definition of **syntax competence levels**,[8] which states

[2]"Visual Interactive Syntax Learning", Institute of Language and Communication (ISK), University of Southern Denmark (SDU) - Odense Campus: https://edu.visl.dk/visl2/

[3]For example, a syntactic labyrinth, as well as a "syntactic Tetris" and other syntactic games were implemented as (now obsolete) java applets.

[4]"Modèles et Traces au service de l'Apprentissage des Langues", Models and learning analytics for language learning.

[5]Learning Management System.

[6]The presented corpus to quiz processing chain is intended for Moodle, but other platforms could be targeted, as long as they allow importing quizzes and exercises from structured text files (XML, json, or other ad-hoc formats).

[7]Other CONLL-U formatted corpora are available for French, but they do not meet the same quality standards as the FTB and Sequoia. Moreover, they are mostly oral transcription corpora, which are not the best material for our purposes.

[8]This definition, inspired by the Common European Framework of Reference for Languages, is still a work in progress,

which syntactic features each learner profile is meant to acquire, from parts-of-speech, to constituent structure and functional relations. Based on the definition of syntax competence levels ranging from A1 (beginner) to C2 (advanced), a set of python scripts have been implemented for the automatic generation of Moodle quiz questions.[9] The scripts process sentences found in a set of CONLL-U corpora according to their overall syntactic types (e.g. simple vs complex syntactic structures, regular vs idiomatic units). Relevant sentences are then transformed into python data structures, and quiz questions are assembled by using each sentence's set of parameters, such as list of words, part-of-speech tag for each word, functions and dependency relations, etc. Different execution parameters allow the user to generate questions for different learner profiles. For example, the instructor can target specific words, lemmas and morphological constraints (e.g. a Noun bearing a *-able* suffix), specific subsets of part-of-speech, or function, tags, or even the number of distractors and syntactic annotation terminology to use (e.g. "SUJET" instead of "nsubj"). Figure 2 shows an example of a GIFT (General Import Format Template) structured quiz question on nouns bearing the *-able* suffix.

```
:: Parties du Discours ::[markdown]   Donner la partie du discours du mot **imputables** dans la phrase:
Très brièvement, il s'agit de limiter les émissions de CO2 **imputables** à l'homme.
{
        ~V_Subj
        ~PROPN
        ~PRO_Int
        ~Conj_de_Sub
        ~DET_Int
        ~V_Part_Prés
        ~Conj_de_Coord
        ~V_Inf
        ~AUX
        =ADJ
}
```

Figure 2: A quiz question on *-able* nouns

In this example, our question-generation script was launched with parameters to target all *-able* nouns. In the generated quiz question, learners must select the proper part-of-speech for noun "imputables" (*attributable*) in the context of the given sentence,[10] extracted from the Sequoia corpus. In this example, 9 distractors are shown (with a minimum of 2). The order in which distractors are presented, and other quiz parameters (e.g. randomized selection of individual questions, total allotted time) are defined by the instructor, for each quiz activity. In this prototype version, feedbacks must be manually provided by the instructor.[11]

Figure 3 shows how Moodle renders a GIFT-structured **part-of-speech quiz**, based on a sentence taken from the French Treebank.[12] Many French native speakers confuse the coordinating conjunction "ou" with the relative pronoun "où". Therefore, we have targeted "ou"/"où" and other typical confusing cases ("et"/"est", etc.) by stating a constraint on the form of the desired lemmas. These exercises are particularly adequate for the first weeks of a syntax course: the sentence is not too long, and the syntactic structure is relatively straightforward (a simple predicative structure). As such, it is adequate for less experienced learners (i.e. A1/A2 syntax competence level). Here, the quiz uses POS-tags available in the CONLL-U formatted version of the FTB, which are adapted and rendered so as to match a syntactic terminology closer to traditional grammar rather than UD categories. This mapping is achieved via a customizable equivalence table, it controls how the syntactic terminology will be presented to the learners, according to their competence level.[13]

In figure 4 a sentence[14] taken from the Sequoia corpus was used to assemble a quiz on **syntactic**

since no widely accepted, explicit definition of syntax competence levels could be found so far.

[9] All CONLL file preprocessing steps are performed thanks to the pyconll library. The code, as well as a large set of GIFT-structured questions are available at https://github.com/abalvet/ACE.

[10] *In a nutshell, the goal is to limit CO2 emissions attributable to mankind.*

[11] Chatbots, such as ChatGPT, might be integrated in future versions, in order to generate feedbacks based on learners's responses and competence levels.

[12] *In the plural form, since this event is economic, ecological, ideological and even iconoclastic.*

[13] Less advanced L1 students can be presented with a rather classical set of grammatical distinctions while more advanced L3 students can be exposed to the terminology and syntactic distinctions following actual UD guidelines.

[14] *On this matter, we ask ourselves the question of why feasibility studies and technical assistance measures amount to 47%*

Figure 3: A parts-of-speech quiz generated from a FTB sentence

functions, by leveraging on the dependency annotations available in the CONLL-U formatted version of the corpus. Here, the expected answer is "COD" (direct object), since *question* is the nominal head of the NP governed by *posons* (ask). As can be seen, the particular question shown is part of a quiz activity comprising 40 questions.

Figure 4: A quiz on syntactic functions generated from a UD corpus

In the examples above, all quiz activities are essentially text files, structured with the GIFT format. As such, the generated questions are ready to import into a Moodle question database to be used either as a formative or as an evaluative assessment activity.

3 First results: automatic syntax quizzes in the classroom

We first introduced automatically-generated syntax quizzes to groups of L1 students enrolled in the linguistics curriculum at our university in 2018. Initially, the aim was essentially to test different Moodle activities, while still retaining a classical "chalk-and-talk" approach. The COVID-19 pandemic, and the lockdowns that followed, have forced us to transform what was initially a mere addition to a classical teaching plan into our main formative and evaluative assessment tool. With a total of over 200 L1 students

of the budget, or nearly 223 million euros.

in 2019-2020, we had to devise a workable corpus-to-quiz solution that would provide large amounts of relevant formative and evaluative activities throughout a whole academic year. We kept using the material developed during that period even after lockdowns were lifted, and we are happy to report that, after having exposed over 800 L1 students to our automatically-generated quizzes over the course of four years, the basic concept can be validated. Students generally find it reassuring to be able to train themselves on large sets of syntax quizzes in preparation for mid-term and end-of-term exams. The fact that Moodle can provide an instant feedback on their performance is a clear motivation and engagement booster, as opposed to traditional syntax exercises. From the instructor's point of view, learner analytics processed by Moodle make it possible to easily identify "hard" or "easy" questions post-hoc, in order to fine-tune our growing set of syntax quizzes. We are now contemplating how to integrate the generated quizzes into other LMS platforms, and how to devise an interactive electronic syntax textbook, by using Jupyter books in conjunction with Moodle and other LMS platforms.

References

Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a Treebank for French. *Treebanks: Building and using parsed corpora*, 165–187.

Bick, E. (2001). The VISL System: Research and applicative aspects of IT-based learning. *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*.

Bick, E. (2004). Grammar for fun: IT-based grammar learning with VISL. *Copenhagen studies in language*, *30*, 49.

Candito, M., Perrier, G., Guillaume, B., Ribeyre, C., Fort, K., Seddah, D., & de La Clergerie, É. V. (2014). Deep Syntax Annotation of the Sequoia French Treebank. *International Conference on Language Resources and Evaluation (LREC)*.

Colin, É. (2020). *Traitement automatique des langues et génération automatique d'exercices de grammaire* (Doctoral dissertation). Université de Lorraine.

De Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, *47*(2), 255–308.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2008). The Sketch Engine. *Practical Lexicography: a reader*, 297–306.

Perez-Beltrachini, L., Gardent, C., & Kruszewski, G. (2012). Generating Grammar Exercises. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 147–156.

Uibo, H., & Bick, E. (2005). Treebank-based research and e-learning of Estonian syntax. *Proceedings of Second Baltic Conference on Human Language Technologies: Second Baltic Conference on Human Language Technologies*, 4–5.

Wijlff, A. (2006). VISL in Danish schools. *English Teaching: Practice & Critique (University of Waikato)*, *5*(1).

Workflows for Semantic Change Research with CLARIN Resource Families

Paola Marongiu
University of Neuchâtel, Switzerland
paola.marongiu@unine.ch

Fahad Khan
Consiglio Nazionale delle Ricerche, Italy
fahad.khan@ilc.cnr.it

Barbara McGillivray
King's College London, United Kingdom
barbara.mcgillivray@kcl.ac.uk

Abstract

Recently, there has been a growing interest and numerous advances in computational and quantitative approaches to research in semantic change (SC). However, the resources to conduct this type of research are currently scattered across different CLARIN Resource Families (CRF) or they are not even included in any CRF. In this paper, we present our preparatory work for the implementation of a new CRF for SC research. The new CRF will host cross-linguistically valid workflows to support SC research in different fields. The goal is to collect relevant existing resources in the CLARIN infrastructure and organise them coherently around macro-areas of research that focus on SC or exploit SC to answer their subject-specific research questions. Firstly, we describe the existing resources and tools for SC and the other CRFs. Then, we outline the structure of the workflows for SC research. Finally, we present an application of the workflows by using the Latin medical lexicon as a case study.

1 Background

Semantic change (SC), the linguistic phenomenon by which words, phrases or expressions change their meanings over time, is of great relevance to humanities and social science research (HSS), with strong connections to concept drift analysis in history, diachronic semantics in historical linguistics, and lexicography. In recent years the NLP community has grown a strong interest in developing computational methods for the automatic detection of SC from corpora, especially at the lexical-semantic level, e.g. the dedicated SemEval task (Schlechtweg et al., 2020) on English, Latin, Swedish and German, and further tasks for Italian, Russian and Spanish (Mickus et al., 2022). In parallel, research on building linked data models to represent and disseminate SC data has led to the development of vocabularies and standards for this task, but a lot remains to be done to make these resources, technologies, and approaches more widely accessible, thus allowing researchers to further advance research in SC change and concept drift. SC research requires several different kinds of language resource and/or tool: annotated corpora, dictionaries/lexica, language models and SC detection algorithms. Many such resources and tools are currently available, but they lay within separate CLARIN resource families (CRFs), as briefly outlined later in this section.

In this presentation we describe the work carried out within the CLARIN-funded project *A new CLARIN Resource Family for lexical semantic change research.*[1].Within this project we have focused on a specific level of semantic change, namely lexical semantic change (LSC), intended as the semantic change of a given lexical item from a semasiological perspective (Koch, 2016, p. 23). The purpose of this project was to lay the groundwork for the creation of a new task-oriented CRF that brings together existing tools and resources needed to support LSC research.

The implementation of a new CRF hosting workflows for LSC research well aligns with CLARIN's 2021-2023 strategy via the following:

[1]The project ran from November 2022 to July 2023 (see McGillivray et al., 2023 for the outputs)

- By enabling multilingual LSC research investigating language as a carrier of cultural content and information.
- By strengthening CLARIN's role as a pillar supporting HSS research given the central role played by LSC in HSS research.
- By enabling LSC detection algorithms on multilingual language resources, facilitating advancement in language technology research.
- By improving discoverability of existing tools and CRFs.

As mentioned above, research in LSC can benefit from already existing resources and tools. Several corpora have been annotated with word senses by using computational lexical resources such as WordNet (WN) (Fellbaum, 1998), BabelNet (Navigli & Ponzetto, 2012) and Wikipedia to build the sense inventory (see the surveys by Petrolito and Bond, 2014 and Pasini and Camacho-Collados, 2020). In many cases these and other resources are available within CLARIN, scattered across different CRFs: the Historical Corpora CRF,[2] the Manually Annotated Corpora CRF[3] and the Dictionaries[4] and Lexica CRF.[5] In addition, new CRFs on language models and Linguistic Linked Open Data in the context of the LiLa project[6] are being planned. What is missing, and what our new CRF proposes, is a way to combine all these existing resources with LSC-specific ones in order to enable HSS and computational linguistics researchers to advance LSC research. Section 2 gives more details of the content of this new CRF.

2 The workflows

The new CRF which we are developing aims at being a cross-transversal resource family which gathers resources scattered across other CRFs, and possibly including new ones. In order to do so, we resorted to developing workflows for LSC research. The idea of workflows was inspired by the model offered by the Social Sciences and Humanities (SSH) Open Marketplace. This portal features various task-oriented workflows designed to guide the users through the resources already available in the portal, enabling them to achieve a specific goal, such as creating a TEI dictionary.[7] A workflow in SSH Open Marketplace breaks down the research process into separate steps, and connects each step with a series of relevant resources (datasets, manuals, digital tools).

Our intention is to introduce workflows within CLARIN, in such a way as to gather CLARIN resources coherently around a specific task (in our case LSC research), but adapted to various different disciplines. The workflows will connect the following main kinds of datasets:

1. *Lexical semantic annotation datasets*. These consist of snippets of diachronic corpora which contain a set of target words and which are annotated at the level of lexical semantics, for example with reference to dictionary definitions or WordNet synsets. An example is the SemEval Latin annotated dataset (McGillivray, 2021; McGillivray et al., 2022) or the Ancient Greek dataset (Vatri et al., 2019). These datasets can serve as training or evaluation sets for lexical semantic change detection algorithms as in Schlechtweg et al. (2020) or they can be used for corpus studies on lexical semantics as in McGillivray et al. (2022).

2. *Trained word embeddings from diachronic corpora*, either *word type* or *word token* embeddings (contextualised embeddings). One example of this kind of dataset is the Latin lemma embeddings (Sprugnoli et al., 2020; Sprugnoli et al., 2021). These datasets are critical to any algorithm that uses distributional information (i.e. data about words' corpus co-occurrence) to derive semantic representations of words for further quantitative processing.

[2] https://www.clarin.eu/resource-families/historical-corpora
[3] https://www.clarin.eu/resource-families/manually-annotated-corpora
[4] https://www.clarin.eu/resource-families/dictionaries
[5] https://www.clarin.eu/resource-families/lexical-resources-lexica
[6] https://lila-erc.eu/
[7] https://marketplace.sshopencloud.eu/workflow/4qFarh

3. *Tools for semantic change detection*. These are algorithms such as those developed in the SemEval 2020 shared task (Schlechtweg et al., 2020).

4. *Computational lexical resources* which provide the sense inventories for annotating the corpora in 1); these can be digitised versions of legacy resources with IDs for individual senses, native born lexical resources or dictionaries, or conceptual resources such as wordnets.[8]

5. *Structured datasets describing lexical etymologies as graphs using standard vocabularies* (Khan, 2018). Although in many cases we may subsume such datasets under the former category, since they are often extracted from legacy lexicographic resources, this is not always necessarily the case. An example is the LiLa representation of the etymological content of de Vaan Etymological Dictionary of Latin as a series of RDF graphs (Mambrini & Passarotti, 2020). There are other digital resources that are currently not available in the CLARIN infrastructure, but would be a valuable part of the workflows. These include the datasets of etymologies provided by the Tower of Babel initiative[9], and tools for extracting, editing and comparing etymologies e.g. the Etymological DICtionary ediTOR (EDICTOR) (List, 2017) and the Etymological Inference Engine (EtInEn)[10] developed by Johannes Dellert and his team.

6. *Other resources*. There are various resources that can be used to determine and describe the type of semantic change observed. Such resources mostly comprise scientific publications rather than digital resources, but there are a few exceptions e.g. The Database of Semantic Shifts (CSSh) (Zalizniak et al., 2012).

The workflows are designed to be applicable in various research fields, as long as they build upon the study of LSC to address their research questions. In this regard, we have proposed workflow drafts for lexicology, lexicography, NLP, history (and sub-fields cultural history and history of ideas) and legal studies. In this article, we present the workflow we have developed for lexicologists aiming to study how words from a specific semantic field have changed over time and/or based on their context of use. To achieve this, we will focus on the Latin medical lexicon as a case study.[11]

3 The workflow for lexicology: a case study on medical Latin

Previous studies have shown that Latin medical terms in this field are often borrowed from everyday language and adopted in the medical lexicon with a specialised meaning acquired through some type of semantic change, such as *metaphor*. Examples are the words *lenticula* from 'lentil' to 'freckle', *mola* from 'mill-stone' to 'molar tooth', *spina* from 'thorn' to 'spine' (Langslow, 2000, pp. 187, 182). For the sake of this case study, we assume that a lexicologist wants to study how words in the Latin medical lexicon changed their meaning from a general to a specialised one. Starting with a specific semantic field and a list of words associated with it, they need to determine which words have changed their meaning and which senses they have acquired or lost, or which of their meanings have undergone some other type of semantic change. The workflow for this research field involves the following steps (including the relevant CRFs for steps 1 to 3):

1. *Set up a corpus*. The researcher will need to perform their analysis on a reference corpus for the target language. The corpus does not necessarily need to be sense-annotated, but it should be lemmatised and PoS tagged, and contain metadata about the texts. The lexicologist will split the corpus into different sub-corpora according to the dimension of variation under study. **CRFs**: Historical Corpora; Manually annotated corpora. In addition to this, the lexicologist might decide to set up a list of words that should be studied in the corpus. The list can be based on previous studies in that

[8] https://www.clarin.eu/resource-families/lexical-resources-conceptual-resources

[9] https://starlingdb.org/descrip.php?lan=en#bases

[10] https://github.com/verenablaschke/etinen-etymology

[11] It should be noted that although this case study targets a specific historical language, the workflows are conceived in order to be cross-linguistically valid. For each step we mention the relevant CRFs, rather than specific items within the CRFs.

semantic field, but also on existing resources within the CLARIN infrastructure. **CRFs**: Glossaries; Lexica; Wordlists.

2. *Annotate the corpus with word senses* (if relevant). If corpora already annotated with word senses are not available, the researcher can manually annotate them. The sense inventory can be built based on lexical resources such as wordnets (see point 4 in section 2). The dataset will also serve as a Gold Standard for the evaluation of the output of computational methods. **CRFs**: Conceptual Resources.

3. *Train word embeddings on sub-corpora and evaluate them*. This helps us determine if and how the semantics of a certain word has shifted and has acquired a new sense in a different domain. The results of the word embeddings analysis will be evaluated against the Gold Standard created in step 2. The researcher compares the closest neighbours of the target words in different time spans in order to determine how the word has changed based on words that occur in similar contexts. **CRFs**: Language models.

4. *Qualify the type of semantic change*. While in CLARIN there is no specific resource family for literature, in the SSH Open Marketplace manuals and specific bibliographic references can be linked to a step of the workflow, as long as they are stored within the infrastructure.

To briefly illustrate how such a workflow is applied to a specific case, let's take the Latin word *spina*. To investigate the type of semantic change observed in this case (from 'thorn' to 'spine'), the lexicologist will first select a reference corpus. In this case, the chosen corpus is LatinISE (McGillivray & Kilgarriff, 2013), which is already part of the CLARIN infrastructure. LatinISE contains a rich set of metadata in addition to PoS and lemmas. The lexicologist divides the corpus into medical and non-medical texts, utilising the metadata set (Step 1). Depending on how they want to structure their research process, the lexicologist might want to have a corpus manually annotated with word senses for medical words. Latin WordNet (LWN, point 4) (Biagetti et al., 2021; Minozzi, 2017) can be used to build the sense inventories. By querying the LWN API it is possible to retrieve all the synsets of the target word *spina*, including n#09422421 'a small spike (as the inflorescence on grasses and sedges)' and n#04330266 'the series of vertebrae forming the axis of the skeleton and protecting the spinal cord' (Step 2). The lexicologist can use word embedding models to determine whether and how the word has changed in meaning from non-specialist to specialist lexicon. To achieve this, they train word embeddings on the two subcorpora created in Step 1 and align them, enabling the evaluation and comparison of results across the two subcorpora. The results for *spina* on the two sub-corpora derived from LatinISE are the following. The 10 nearest neighbours for the model trained on the non-specialised subcorpus contain words such as *vepres* 'thornbush' (0.752), *stramentum* 'corn-stalk' (0.728), *virgultum* 'thornbush' (0.719), and in general words that refer to entities that are characterised by the presence of thorns or spikes. Vice versa, in the model trained on the medical subcorpus the 10 nearest neighbours contain *vertebra* 'vertebra' (0.998), *sinuo* 'to bend in the shape of an arc' (0.998), *costa* 'rib' (0.997) (Step 3) (Marongiu & McGillivray, 2023). To conclude, the lexicologist can ascertain that there is indeed a semantic change affecting the word *spina*, which becomes specialised in the medical domain through a metaphorical shift based on the physical resemblance between the two referents (Langslow, 2000, pp. 181–2) (Step 4).

In this work, and through the example of the word *spina*, we showcase the structure and practical applications of the new CRF that we propose. Specifically, we demonstrate that the implementation of workflows for LSC research in the CLARIN infrastructure can enhance the impact of other existing CRFs, and provide a way to effectively combine them in a real research process.

Acknowledgments

This work was funded by CLARIN-ERIC through the CLARIN Resource Families project. We also wish to thank the experts Marton Ribary, Sandeep Soni, Lauren Klein, Jacob Eisenstein, Dani Roytburg and Emily Bell who provided their feedback on the workflows concerning their fields of expertise.

References

Biagetti, E., Zanchi, C., & Short, W. M. (2021). Toward the creation of WordNets for ancient Indo-European languages. *Proceedings of the 11th Global Wordnet Conference*, 258–266.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. MIT Press.

Khan, A. F. (2018). Towards the Representation of Etymological Data on the Semantic Web. *Information*, *9*(12), 304. https://doi.org/10.3390/info9120304

Koch, P. (2016). Meaning change and semantic shifts. In P. Juvonen & M. Koptjevskaja-Tamm (Eds.), *The lexical typology of semantic shifts* (pp. 21–26). Walter de Gruyter GmbH & Co KG.

Langslow, D. R. (2000). *Medical latin in the roman empire*. Oxford University Press.

List, J.-M. (2017). A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 9–12.

Mambrini, F., & Passarotti, M. (2020). The Etymological Dictionary of Latin and the other Italic languages in LiLa (EDLIL) [ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa].

Marongiu, P., & McGillivray, B. (2023). Lexical semantic change analysis in latin: A use case on medical latin. *Digital Classicist London seminar 2023*.

McGillivray, B. (2021). Latin lexical semantic annotation. https://doi.org/10.18742/16974823.v1

McGillivray, B., Khan, F., & Marongiu, P. (2023). A new CLARIN Resource Family for lexical semantic change. Final report. https://doi.org/10.5281/zenodo.8156200

McGillivray, B., & Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin. In P. Bennett, M. Durrell, S. Scheible, & R. J. Whitt (Eds.), *New methods in historical corpus linguistics* (pp. 247–257). Narr.

McGillivray, B., Kondakova, D., Burman, A., Dell'Oro, F., Bermúdez Sabel, H., Marongiu, P., & Márquez Cruz, M. (2022). A new corpus annotation framework for latin diachronic lexical semantics. *Journal of Latin Linguistics*, *21*(1), 47–105. https://doi.org/https://doi.org/10.1515/joll-2022-2007

Mickus, T., Van Deemter, K., Constant, M., & Paperno, D. (2022). SemEval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 1–14. https://doi.org/10.18653/v1/2022.semeval-1.1

Minozzi, S. (2017). Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. In P. Mastandrea (Ed.), *Strumenti digitali e collaborativi per le scienze dell'antichita* (pp. 123–134). Università Ca' Foscari.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.

Pasini, T., & Camacho-Collados, J. (2020). A short survey on sense-annotated corpora. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5759–5765.

Petrolito, T., & Bond, F. (2014). A survey of wordnet annotated corpora. *Proceedings of the Seventh Global WordNet Conference*, 236–245.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 task 1: Unsupervised lexical semantic change detection. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1–23. https://doi.org/10.18653/v1/2020.semeval-1.1

Sprugnoli, R., Moretti, G., & Passarotti, M. (2020). Building and comparing lemma embeddings for Latin. Classical Latin versus Thomas Aquinas. *Italian Journal of Computational Linguistics*, *6*(1), 29–45.

Sprugnoli, R., Passarotti, M., & Moretti, G. (2021). Lemma embeddings for latin. http://hdl.handle.net/20.500.11752/OPEN-996

Vatri, A., Lähteenoja, V., & McGillivray, B. (2019). Ancient greek semantic change - annotated datasets and code. https://doi.org/10.6084/m9.figshare.c.4445420

Zalizniak, A. A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak, T., & Russo, M. (2012). The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, *50*(3), 633–669.

Standards Information System for CLARIN centres and beyond

Piotr Banski
Leibniz Institute for the German Language
Mannheim, Germany
banski@ids-mannheim.de

Eliza Margaretha Illig
Leibniz Institute for the German Language
Mannheim, Germany
margaretha@ids-mannheim.de

Abstract

In the present contribution, we describe the features of the CLARIN SIS (Standards Information System) that have been designed to assist data-deposition centres in CLARIN. We also show what is needed to go beyond the originally designated target, in order to provide service to sibling and descendant research infrastructures, of which DARIAH and Text+ are taken as examples.

1 Introduction

Many modern research infrastructures offer data deposition services for their users. For CLARIN B-centres, the provision of this service is a default characteristic that is subject to certification requirements and that is used as a basis of a measurement needed to calculate one of the CLARIN-ERIC Key Performance Indicators (see (Bański and Hedeland, 2022) for discussion and further references).

CLARIN is not the only research infrastructure focussing on language resources. CLARIN's focus has historically overlapped with some areas served by DARIAH and, by a natural extension, with CLARIAH networks that combined DARIAH and CLARIN nodes in some of the European countries, at various points in time. In Germany, the national CLARIN-D merged with DARIAH-DE into CLARIAH-DE in 2019, and, since 2022, many former German DARIAH and all the former CLARIN-D centres (as well as some centres previously not belonging to either of the two) have formed the Text+ consortium (part of the German National Research Data Infrastructure, NFDI).

This is illustrated in Figure 1 below, which does not take historical developments into account, but is rather meant to hint at the resulting relationships. The reader should bear in mind that, while CLARIN and DARIAH are multinational networks, Text+ is restricted to Germany.

In the present contribution, we place the Standards Information System, originally conceived within CLARIN (and, to be precise, contributed to the infrastructure by CLARIN-D) in the context of the extended network of inter-RI relationships.

2 Standards Information System: basic information

The current CLARIN Standards Information System[1] extends the former CLARIN Standards Guidance (Stührenberg et al., 2012). Originally, the system was intended to help users in the task of selecting standards most appropriate for their purpose. It attempted to achieve that by providing information about various standards and specifications, and presenting relations between them. A side goal was to provide a taxonomy or a knowledge base of standards and specifications, served by eXist-

[1] The SIS can be accessed at https://standards.clarin.eu/sis/ , which is an alias for https://clarin.ids-mannheim.de/standards/ . Its Github home is at https://github.com/clarin-eric/standards and the documentation is in the project wiki at https://github.com/clarin-eric/standards/wiki . The SIS is listed as a knowledge base at Fairsharing.org: https://fairsharing.org/4705 .

DB (Siegel & Ritter, 2014). That part of the functionality is still a subset of the current SIS, albeit somewhat dated due to limited resources that are needed to maintain and extend the knowledge base.

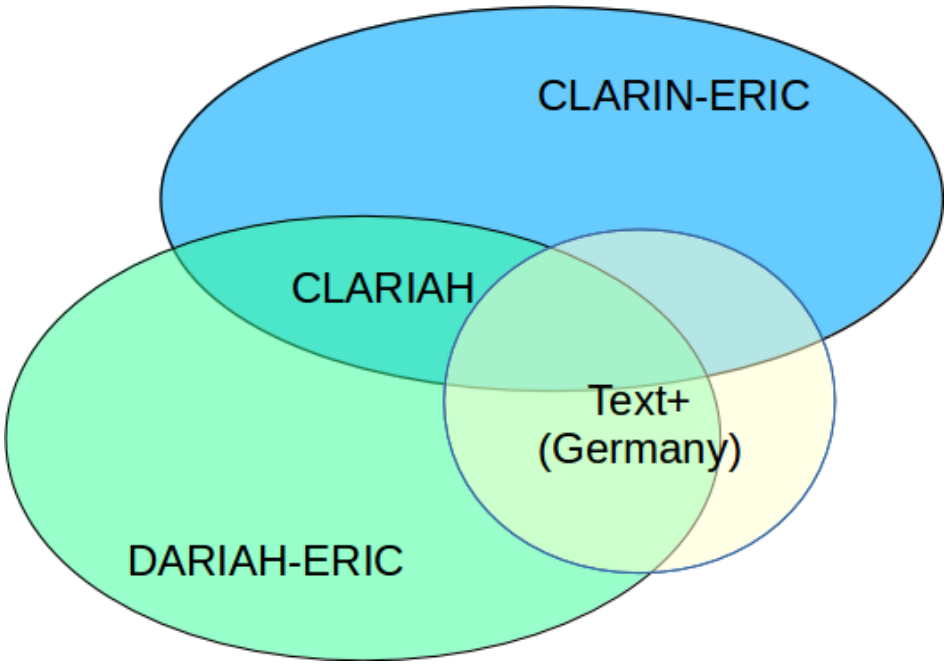


Figure 1: Relationships between language-oriented infrastructures that the SIS will be able to serve in the next step of its development.

Around the year 2019, the CLARIN Standards Committee undertook a task of providing a platform for centres to share their recommendations concerning the formats for data that could be deposited at each of them. Figure 2 (adapted from Bański and Hedeland, 2022), shows the extension to the original data model needed to address that new task.

The task faces several challenges, among others concerning the way to collect the data-deposition formats for each centre and to provide an easy way for the centres to submit their recommendations. Due to the complexity of how data formats are used in HLT research, it is also a challenge to represent the recommendations and maintain them with minimal effort. The deliverable has evolved since 2021 from a complex set of spreadsheets that put together formats, format categories and CLARIN centres, to the current XML format integrated in the SIS.[2]

Figure 2 illustrates, among others, the structure of data-deposition format recommendations: a recommendation is a pairing of a format with a functional domain, accompanied by one of the three recommendation labels: “recommended”, “acceptable”, and “discouraged”.

A crucial element of the system is the set of functional data domains that serve to fine-tune the purposes for which the individual data items are collected: for example, data coming in the PDF/A format are perfect for the purpose of documentation, but definitely not ideal for the purpose of providing annotation for audiovisual sources, or collections of statistical data. This is illustrated in Figure 3.

3 Current SIS functionality for CLARIN centres

The current offer of the SIS towards centres can be summed up in the following three points:

1. increasingly user-oriented way of submitting information
2. increasingly attractive way to benefit from data aggregation
3. a way to reuse the data submitted by the centres

[2] Much of the history behind the task described here is documented at https://www.clarin.eu/content/standards .

Below, we elaborate on each of these three points.

3.1. The preferred way for data submission is by pull requests directed at the SIS source code deposited on GitHub. CLARIN developers are familiar with GitHub, so submitting a PR presents no obstacle. For less technical users, the SIS offers an alternative way through editing the recommendation documents, which may be exported from the section of the SIS devoted to the given centre (even if the set of recommendations is empty). These exported files contain placeholders and templates, in order to make the data input easier. They are additionally constrained by document grammars (W3C XML Schema and ISO Schematron), which signal errors and provide closed lists of options to choose from, where feasible. Finally, many places in the SIS offer an option to switch to editing a templated GitHub feature request, in a single click. This final way is naturally best used for minor fixes. The wiki system that accompanies the SIS source, linked from the SIS instance, provides additional instructions.

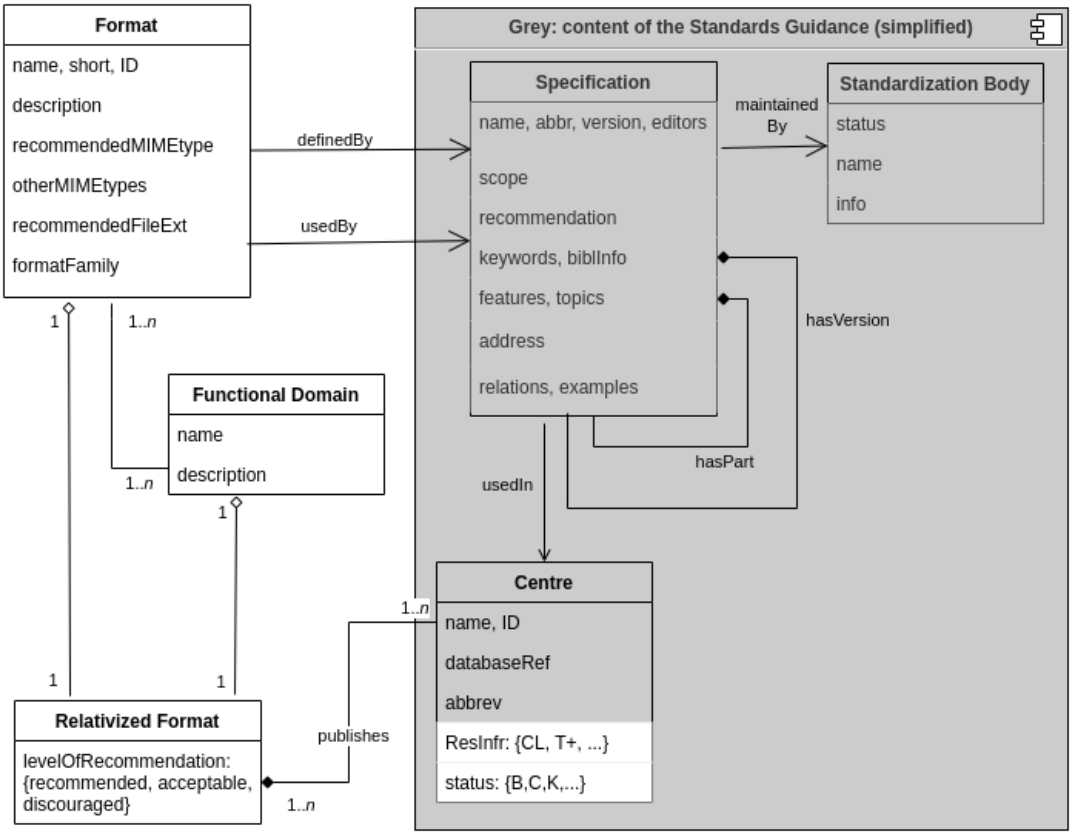


Figure 2: Simplified data model (on grey background) with additions designed to incorporate format recommendations and research infrastructures other than CLARIN. Filled diamond arrows represent strong aggregation, hollow diamond arrows – weak aggregation, while simple relationships are represented by lines, with simple arrowheads pointing to the objects of the relationships.

3.2. Aggregating structured data from several sources presents an opportunity to visualise the data in various ways, and to provide statistics. For this purpose, the SIS offers, among others, word-clouds based on the format keywords, tabular displays of various sorts, extracted lists of file extensions and media types for use in processing pipelines, as well as higher-level statistics concerning, for example, the most “popular” file formats, relative to the intended function of the submitted data. For CLARIN, the data aggregated in the SIS make it possible to dynamically compute one of the Key Performance Indicators (KPIs).

3.3. Finally, the SIS offers a way for the centres to reuse the data that was submitted, via a REST API. This way, the SIS may be used as the sole tool for the maintenance of centre recommendations (and, in the case of CLARIN, to satisfy the B-centre certification requirements). There is no need to manage two separate instances of data: one for the SIS, and one for the centre itself to display. The API offers a way to receive the information that the centre has provided, to be transformed and styled in the way that the centre wishes.

Format	Centre	Domain	Recommendation
PDF/A	EKUT	Image Source Language Data	recommended
PDF/A	EKUT	Textual Source Language Data	recommended
PDF/A	FIN-CLARIN	Documentation	recommended
PDF/A	FIN-CLARIN	Textual Source Language Data	acceptable
PDF/A	IDS	Documentation	recommended
PDF/A	IDS	Image Source Language Data	recommended
PDF/A	DANS	Documentation	recommended
PDF/A	DANS	Other	acceptable
PDF/A	DANS	Textual Source Language Data	recommended
PDF/A	MI	Image Source Language Data	recommended
PDF/A	MI	Textual Source Language Data	recommended
PDF/A	ZIM	Image Source Language Data	recommended
PDF/A	ZIM	Textual Source Language Data	recommended
PDF/A	LAC	Contextual Data	recommended

Figure 3: Fragment of format recommendations by CLARIN centres concerning the PDF/A format. Centres may comment on their recommendations (the circled *i* shows the comment in a pop-up).

4 Extending the SIS beyond CLARIN

The SIS is in the process of constant development and receives upgrades of functionality on a nearly weekly basis. The most recent work has been influenced by meetings with the Text+ Standardisation Group of the Collections cluster, and resulted in partial internationalisation of the underlying functionality: it is now possible to use language tags for centre descriptions and comments on recommendations, and to retrieve that information via the SIS API.[3]

As for the needs of the sibling infrastructure DARIAH, including the cases where the national CLARIN and DARIAH nodes operate as CLARIAH, the SIS offers functional inventory that goes beyond pure language-oriented applications.[4] Depending on the decision by the DARIAH governance (or by the individual repositories) to use the SIS, it remains to be seen whether the repertoire currently offered is going to require further adjustments and fine-tuning given the needs of DARIAH centres.

The SIS also provides a function that enables users to easily switch between RI environments and to filter the web-content by their RIs, so that only information to the selected RI is presented. For instance, when the Text+ environment is selected, only Text+ centres and their format recommendations are listed, whilst those of other centres are hidden. Moreover, language preferences are also taken into account in RI environments. For Text+, which prefers the German language, descriptions and comments are shown in German, as long as centres have provided them (otherwise, the system falls back to English).

Extending the SIS beyond CLARIN opens new challenges and exposes some limitations of the system. First, some CLARIN centres may appear under different names in research infrastructures

[3] See the example result of an API query for the data of IDS Mannheim at https://clarin.ids-mannheim.de/standards/rest/data/recommendations/IDS-recommendation.xml . The API also supports searching and exporting recommendations with some filtering criteria, such as centre, domain and recommendation level.
[4] See https://clarin.ids-mannheim.de/standards/views/list-domains.xq

other than CLARIN. Currently, the system only allows a single name for a single centre. Whether this is acceptable or whether the centre list needs to be split depending on the RI remains to be seen.

Second, since format recommendations are defined per-centre, they are considered to be the same across the RIs. When a centre contributes to multiple RIs, SIS assumes that the format recommendations are shared in these RIs. That means that it would not be possible for the IDS, for example, to recommend the CHAT format in CLARIN but discourage it in Text+. Whether this restriction is going to be problematic, remains to be seen when more centres have provided their data.

5 Summary

The Standards Information System is a dynamic platform that adjusts to the expanding demands of data deposition centres. It used to be a relatively static information booth, which around the year 2020 began to evolve into a partially interactive system. The year 2023 is another road marker on its path, as the system opens towards infrastructures other than CLARIN-ERIC.

6 Acknowledgements

The SIS has been developed in the context of the work done by the CLARIN Standards and Interoperability Committee and owes much to its former and present members, as is only partially evidenced in the CLARIN Bazaar presentations that we offered in the previous years – a lot of ideas have been discussed, criticised and advanced during the (mostly virtual) committee meetings. We would also like to acknowledge the three anonymous CLARIN reviewers and thank them for kind words and critical remarks.

Consortia and Infrastructures mentioned above

CLARIAH-DE: https://www.clariah.de/en/

CLARIN: https://www.clarin.eu/

CLARIN-D: https://clarin-d.net/en/

DARIAH: https://www.dariah.eu/

DARIAH-DE: https://de.dariah.eu/

NFDI: https://www.nfdi.de/

Text+: https://www.text-plus.org/en/home/

References

Bański, P. & Hedeland, H. (2022). Standards in CLARIN. In *CLARIN: The Infrastructure for Language Resources*, Fišer, D. & Witt, A. (eds). Berlin, Boston: De Gruyter, 2022, pp. 307-340. https://doi.org/10.1515/9783110767377-012 (accessed on September 11, 2023)

Siegel, E. & Retter, A. (2014). *eXist*. O'Reilly Media, Inc. ISBN: 9781449337100

Stührenberg, M. & Werthmann, A. & Witt, A. (2012). Guidance through the standards jungle for linguistic resources. In Proceedings of the LREC-12 workshop on collaborative resource development and delivery. Istanbul, Turkey, May 2012, 9–13. European Language Resources Association (ELRA). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/4494/file/Stuehrenberg_Werthmann_Witt_Guidance_through_the_standards_jungle_2012.pdf (accessed on September 11, 2023)

The CLARIN:EL infrastructure

Maria Gavriilidou
ILSP / Athena RC, Greece
maria@athenarc.gr

Stelios Piperidis
ILSP / Athena RC, Greece
spip@athenarc.gr

Dimitrios Galanis
ILSP / Athena RC, Greece
galanisd@athenarc.gr

Juli Bakagianni
ILSP / Athena RC, Greece
julibak@athenarc.gr

Penny Labropoulou
ILSP / Athena RC, Greece
penny@athenarc.gr

Athanasia Kolovou
ILSP / Athena RC, Greece
akolovou@athenarc.gr

Dimitris Gkoumas
ILSP / Athena RC, Greece
dgkoumas@athenarc.gr

Miltos Deligiannis
ILSP / Athena RC, Greece
mdel@athenarc.gr

Kanella Pouli
ILSP / Athena RC, Greece
kanella@athenarc.gr

Iro Tsiouli
ILSP / Athena RC, Greece
tsiouli@athenarc.gr

Leon Voukoutis
ILSP / Athena RC, Greece
leon.voukoutis@athenarc.gr

Katerina Gkirtzou
ILSP / Athena RC, Greece
katerina.gkirtzou@athenarc.gr

Abstract

This paper presents the CLARIN:EL infrastructure, which comprises three pillars: the language resources and technologies Platform, the Portal and the Knowledge Centre. It serves as a comprehensive and interoperable environment that supports language-related research in the fields of language technology, language studies, digital humanities, and political and social sciences. The Platform facilitates language resources sharing by providers, and access to these resources by consumers. The Portal and the K-Centre offer complementary informative material and support services to the community, including awareness raising and training activities. This paper discusses the CLARIN:EL architecture, its design and implementation principles, the functionalities offered to the users, the support activities provided, and the network that enables its operation.

1 Introduction

CLARIN:EL is the Greek National Infrastructure for Language Resources & Technologies (LRTs), which comprises three interconnected pillars, namely, the Platform, the Portal and the NLP:EL Knowledge Centre. CLARIN:EL serves as a comprehensive and interoperable environment that supports language-related research in various fields, such as language technology (LT), linguistics, language studies, digital humanities (DH), political and social sciences. The Platform hosts the LRTs and provides the user interaction mechanisms through appropriate interfaces. The Portal and the K-Centre offer informative material and support the community as regards awareness, training, and knowledge transfer in Language Technology (LT) and Digital Humanities (DH).

The CLARIN:EL network supporting the Infrastructure consists of 14 organization members (9 Universities and 5 Research Centres). Anyone (academics, researchers, students, or the general public), whether affiliated to a network member organization or not, can have full access rights to the infrastructure. Registered users, authenticated via their academic or personal accounts, can upload their resources and/or tools, use the available resources and process them using the services offered by CLARIN:EL.

Resources provided by network members are associated, through the relevant metadata, to the specific organization, while those provided by individuals non-affiliated to a member organization are connected to the Hosted Resources Repository.

2 The CLARIN:EL Platform

CLARIN:EL is part of the Greek Roadmap for Research Infrastructures; currently, it forms part of the APOLLONIS infrastructure, together with DARIAH/DYAS. It supports the community through a certified CLARIN B-Centre and a K-Centre, it has been awarded the CoreTrustSeal, and it is listed in re3data.org.

The CLARIN:EL Platform consists of two interconnected subsystems: (a) a system for documenting (through the Metadata editor), and for storing, sharing, searching, retrieving, and downloading language resources (through the Central Inventory). It currently contains 802 resources (659 corpora, 92 lexical resources, 49 tools/services, and 2 language descriptions); and (b) the CLARIN:EL workspace: a system providing integrated services that perform core Natural Language Processing (NLP) tasks, i.e., sentence splitting, tokenization, PoS tagging, lemmatization, parsing, chunking, named entity recognition, as well as other tasks such as text classification and verbal aggression analysis. Moreover, it offers services that perform data format and character encoding conversion. The CLARIN:EL Central Inventory aims at providing a comprehensive overview of existing resources for Greek (alone or in combination with other languages). This includes listing metadata records and their CLARIN:EL hosted data or software, as well as metadata records whose corresponding data or software reside outside the CLARIN:EL platform, but also metadata records with no proper datasets (i.e., bibliographical lists, useful catalogs, etc.).

2.1 Documentation of resources with metadata

To ensure appropriate description of deposited resources, CLARIN:EL provides a rich metadata schema, CLARIN-SHARE, which allows coherent documentation to be added to each resource. The CLARIN-SHARE metadata model builds upon the META-SHARE metadata model (Gavriilidou et al., 2012), and its application profiles, ELG-SHARE (Labropoulou et al., 2022), ELRC-SHARE (Piperidis et al., 2018), and the MS-OWL ontology (Khan et al., 2022; McCrae et al., 2015), RDF/OWL representation of the model. The schema is intertwined with and supports the full lifecycle of language resources, from creation to annotation and usage. To foster the visibility and reusability of data, CLARIN:EL exposes metadata for harvesting, thus extending their discovery. The CLARIN-SHARE metadata schema has been converted into broadly used metadata schemas, such as CMDI, DC and OLAC, and the metadata records of the resources are harvested by repositories and infrastructures that support such metadata schemas (e.g., CLARIN Virtual Language Observatory/VLO).

Resource providers in CLARIN:EL have two options for creating metadata for their resources: create and upload XML files that adhere to the CLARIN-SHARE metadata schema or create metadata records using the platform's metadata editor. The submitted XML files are automatically checked for completeness and well-formedness. The metadata editor guides the providers to the complete description and uploading of their resources, it safeguards interoperability by using controlled vocabularies (where applicable) and assists them with examples and tips. The completion of the description and the automatic checking are followed by two rounds of manual assessment. The first round involves metadata and legal validation performed by human validators, followed by the final approval by the supervisor of each organization member, which triggers the resource's publication in the repository. Additionally, frequent quality checks, aiming at the completeness and correctness of metadata records and related datasets, are conducted centrally by the dedicated CLARIN:EL technical and metadata team.

2.2 Deposition of Language Resources

Data providers can get guidance and assistance via the Help pages[1] and the relevant Policy documents[2],[3]. The repository offers support on various issues such as data formats, metadata, and legal aspects, through

[1] CLARIN:EL User manual https://clarin-platform-documentation.readthedocs.io/en/stable/
[2] Data Collection policy: https://www.clarin.gr/sites/default/files/CLARINELDataCollectionPolicy.pdf
[3] Deposition documentation: https://clarin-platform-documentation.readthedocs.io/en/stable/all/4_Data/DataPreparation.html?highlight=deposit

the Recommended Formats guidelines, online documentation for metadata and data preparation, documentation and deposition, video tutorials, and helpdesks.

Resources deposited encompass written, spoken, or multimodal content. They can be texts, lexical resources, language models or processing tools, and they may pertain to modern Greek language, to earlier forms of Greek, or to other languages. To be processable by the integrated services of CLARIN:EL, the resources must be in one of the recommended text formats (plain text, XML, TMX, etc.).

CLARIN:EL favors and promotes Open Licenses; however, distribution and/or use restrictions on data are respected. Metadata are freely accessible to all with a CC-BY 4.0 license. The responsibility of clearing IPR and selecting the appropriate license lies with the resource provider. CLARIN:EL offers a variety of standard licenses for the provider to select from, and assistance through the Legal Helpdesk.

2.3 Searching and retrieval of Language Resources

Through the platform, users can search for resources using keywords and facets, or browse the resource catalogue and select a resource to view its full description; if interested, they can download it, or use the NLP services of CLARIN:EL to process it. CLARIN:EL presents resources in one central inventory. The catalogue lists metadata records (with or without data). Resources with no data fall into two categories: (a) metadata records in anticipation of the data that is not yet ready to be published and (b) meta-resources, i.e., ancillary resources (e.g., bibliographical lists, literature reviews, etc.). Browsing, viewing, and exporting metadata records, as well as downloading open-access resources are available to all users (registered or not), while user authentication and authorization are required for using the CLARIN:EL processing services or accessing restricted resources. The downloadability of a resource depends on the license defined by the provider. Legal and technical restrictions on resources are specified by the provider via the relevant metadata elements, based on which CLARIN:EL implements the resource's access policy. For the content files of a resource to be accessible, two criteria must be met: an open access license, and storage of the content files at an access point within CLARIN:EL or externally.

2.4 Processing of Language Resources

CLARIN:EL offers two types of tools/services for processing data: (a) single-task tools (e.g., lemmatizers, tokenizers, etc.) available as web services or as downloadable tools, accessible either from within CLARIN:EL or through an external link, and (b) the Workspace, which includes NLP web services integrated in the CLARIN:EL infrastructure. Each single-task web service can also be part of a workflow, i.e., of a pipeline of tools that operate at multiple levels of analysis (e.g., a workflow starting from sentence splitting, continuing with tokenization, POS tagging, lemmatization and concluding with named entity recognition). The Workspace is designed to support non-expert users in their data processing tasks, by providing ready-to-use pipelines of interoperable tools at a single click, thereby relieving them from the burden of selecting, downloading, and assembling tools from scratch. Users can process datasets hosted at CLARIN:EL or upload and process their own datasets (with a size limit of 2MB). In the former case, the processing results are stored in the infrastructure as a new resource, with its metadata automatically generated combining the metadata of the dataset and of the processing service used. The outcome of the processing is available both in the data format generated by the last service of the workflow (such as XML or XMI), and also in Comma Separated Values (CSV) format. The latter is provided for reasons of user-friendliness and interoperability (such files can be fed to other NLP services or to visualization tools).

2.5 User and Resource-lifecycle management

Registered users have full access to all CLARIN:EL functionalities and are considered potential resource providers, either as individuals or as members of their organization. The activities available to the users depend on their roles, which are defined in the User Management module (based on Keycloak). The User Roles schema comprises the roles of Curator (assigned to all registered users), Validator (assigned by the Supervisor), and Supervisor. These roles are involved in the creation and publishing of a resource, with varying rights: Supervisors have the full list of permitted actions, Validators are responsible for the legal and metadata quality check, while Curators have the basic set of actions.

The set of states of a resource in the process of being prepared for publication in the Central Inventory is depicted in the Resource Lifecycle; these states include the creation of a new resource by a curator (resource status: Draft), the automatic checking of its syntactic validity and conformity with the specifications (status: Ingested/Syntactically valid), the submission of the resource by the curator and the assignment of the resource to validators by the supervisor (status: Assigned for Validation); after the approval of the resource by the validators (status: Approved), the supervisor publishes the resource, making it visible on the CLARIN:EL inventory (status: Published).
Each member organization is responsible for its internal User Role Management, i.e. assigning roles (Curator, Validator, Supervisor), and for ensuring efficient creation, description, and publication of their own resources. Above this User Role Management at the level of member organizations, additional Validator and Supervisor roles exist at the central level of the CLARIN:EL Platform, with rights on all resources, facilitating quality assessment to ensure completeness and correctness.

3 The CLARIN:EL technical architecture

All the above functionalities are supported by the CLARIN:EL architecture, designed with state-of-the art technologies. Its subsystems are built with robust, open-source, scalable technologies, and consist of several applications: the PostgreSQL database (DB) used for storing several types of data, such as user data, the metadata records of the LRs, etc.; ElasticSearch for indexing; the repository backend, built using the Django web framework, offers REST services for managing metadata (import, create, update, delete), authorizes access to the resources etc.; the repository, based on the META-SHARE software[4], with many improved architectural choices, new functionalities and features; the User Interface that consists of web pages for searching/browsing the catalogue, the metadata editor for creating/updating metadata, admin pages for validating resources etc.; Keycloak, an identity and access management solution used for securing the applications; the integrated NLP services; a manager responsible for executing NLP services and a scheduler that decides where/when a user's processing request will be executed, to avoid platform overloading; and finally the User Dashboard.

The CLARIN:EL User Dashboard, available only to registered users, is a Single Page Application (SPA), built using React, providing users with a quick and easy way to monitor and track the performance of their tasks while interacting with various CLARIN:EL resources and services. The main features of the dashboard include customization (different dashboard for each user role), interactivity, real-time data display, alerts, and notifications. The User Dashboard serves both as an entry point to create and upload resources and use NLP processing services, and also as an overview page presenting the users' activity history (information on the resources created, tasks and processing jobs), as well as their editable profile.

All the above applications run as Docker containers at a Kubernetes (k8s) cluster, maintained and supported by the CLARIN:EL development team in ILSP/Athena RC. The checked LRs data are saved in a dedicated Network Attached Storage, while metadata are stored in PostgreSQL. CLARIN:EL uses Handle.net service to assign PIDs to resources, to ensure data accessibility. Procedures are in place for ensuring that hardware, software, and storage media containing archival copies of digital content are managed in accordance with security control, data protection and recovery standards.

4 User Support

CLARIN:EL provides several assistance mechanisms to support user needs. The Portal includes (i) information material on the infrastructure, the use of the Platform and the provided services, FAQs, etc., (ii) dissemination material (news, events, etc.), and (iii) educational material (video tutorials, scientific publications, and presentations). Publicly accessible Helpdesks enable interested parties to ask questions on technical, documentation and legal issues. The Portal, besides hosting the informative material mentioned above, also provides links to redirect the users to the Platform and the NLP:EL Knowledge Centre.

The K-Centre, which aims at actively supporting research and scientific advances in the relevant fields, is organized in two main units; *Knowledge*, where users can find LT tools and services, information on studies and curricula, educational and training material regarding NLP, and *Community*,

[4] https://github.com/metashare/META-SHARE

where they can be informed on NLP/LT teams in Greece, certified CLARIN K-Centres and National and European LRTs Infrastructures.

CLARIN:EL also provides detailed online documentation on the Platform and all its functionalities. The User Manual familiarizes the users (provider, curator or consumer) with the basic concepts of the Infrastructure, guides them through its main functionalities (browsing, searching, viewing, downloading, and processing Language Resources), instructs them how to create and manage their resources, and explains the role and the significance of the metadata schema used for this purpose. Finally, it provides crucial information on legal issues connected to the publication, distribution, and use of language resources (licensing), as well as those connected to the use of the infrastructure itself (Privacy policy and Terms of Use).

In addition to the management and the continuous updating of the material provided through the Portal and the NLP:EL Knowledge Centre, the CLARIN:EL team organizes training activities, such as webinars, workshops, summer schools, datathons, etc., (single or recurrent) in order to educate users on Language Technology and Digital Humanities, to raise awareness or to introduce new functionalities of the Platform.

5 Conclusion

We have presented the CLARIN:EL infrastructure, the functionalities available to the users, the design and implementation principles as reflected in its architecture, and the support activities provided to the community. Future steps include the maintenance and upgrading of the infrastructure's modules, the population of the repository with new resources, including workflows, the continuous support of its users, the enlargement of the network with new organization members and end-users, the interoperability with other infrastructures and repositories, and, finally, the hosting of outreach activities aiming to raise awareness about LT in the research community.

Acknowledgements

This work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.), under the Emblematic Action "The emerging landscape of digital work in Humanities in the context of the European infrastructures DARIAH and CLARIN" (Project Number: 7982), https://digital-landscape.gr/.

References

Gavriilidou, M., et al. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1090-1097, http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf

Khan, A.F., et al. (2022). When linguistics meets web technologies. Recent advances in modelling linguistic linked data. Semantic Web Preprint (2022): 1-64. https://www.semantic-web-journal.net/content/when-linguistics-meets-web-technologies-recent-advances-modelling-linguistic-linked-open

Labropoulou, P., et al. (2022). Making metadata fit for next generation language technology platforms: The metadata schema of the European Language Grid. arXiv preprint arXiv:2003.13236

McCrae, J.P., et al (2015). One Ontology to Bind Them All: The META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. *The Semantic Web: ESWC 2015 Satellite Events*. ESWC 2015. Lecture Notes in Computer Science, vol 9341. Springer, Cham. https://doi.org/10.1007/978-3-319-25639-9_42

Piperidis, S., et al. (2018). Managing public sector data for multilingual applications development. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018*), Miyazaki, Japan. European Language Resources Association (ELRA). https://aclanthology.org/L18-1205.pdf

NB DH-LAB: a corpus infrastructure for social sciences and humanities computing

Magnus Breder Birkenes
National Library of Norway
magnus.birkenes@nb.no

Lars Johnsen
National Library of Norway
lars.johnsen@nb.no

Andre Kåsen
National Library of Norway
andre.kasen@nb.no

Abstract

The paper introduces NB DH-LAB, a corpus infrastructure for the social sciences and humanities computing, developed at the National Library of Norway (NLN), a CLARIN C-centre. Having digitized nearly the complete written published cultural heritage of Norway, NLN has built a corpus infrastructure adhering to the FAIR principles. The paper discusses the various building blocks of the infrastructure and shows some basic examples of its usage.

1 Background

In 2006, the National Library of Norway (henceforth NLN) embarked on an ambitious digitization programme. The goal was to digitize all its collections in the forthcoming years. The collections contain all material collected under the Norwegian Legal Deposit Act, such as books, newspapers, journals, music, movies, posters and maps, practically everything published in the public domain in Norway during the last 500 years. As of 2023, practically all books ever published in Norway (albeit with certain exceptions) and most newspapers have been digitized. At the moment, newspapers and journals are being processed. Likewise, other primarily non-textual collections like movies, photos and broadcasting material are being digitized at a rapid scale. We will focus on the written text sources here.

Digitization is of limited value for a national library if access to the content is massively restricted: Our mission is to make the cultural heritage available. In Norway, the so-called Bokhylla agreement from 2012[1] grants all users in Norway, i.e. users with a Norwegian IP address, access to books published before the year 2001. Furthermore, educational staff at Norwegian universities enjoys full access to vast amounts of the digitized books, newspapers and journals for research purposes. The objects can be browsed using the online library *Nettbiblioteket*[2], a IIIF based viewer, with full-text search powered by Elastic.

In 2014, the NLN created NB N-gram (cf. Birkenes et al., 2015), a trend viewer similar to Google Books Ngram Viewer (cf. Michel et al., 2011), so that researchers and the general public can explore our collections quantitatively, regardless of access to the full-text.[3] To this aim, we extracted unigrams, bigrams and trigrams from all books and newspapers digitized at that time and aggregated them for each year and language in the metadata. The dataset[4] and service have since been updated as the corpus has continued to grow.

NB N-gram is a good starting point for corpus exploration and hypothesis generation, but we soon noticed that researchers wanted to dig deeper into the texts. In 2016, the NLN started to give researchers restricted access to its corpora via a corpus platform providing frequency lists, concordances and collocations. The leading idea was to give researchers access to as much text and aggregations over text as possible, without violating copyright law. This corpus infrastructure, which we simply call NB DH-LAB (i.e. National Library of Norway Digital Humanities Laboratory), is described in this paper.

[1] https://www.kopinor.no/avtaletekster/bokhylla-avtalen (last visited: 2023/04/13)

[2] https://www.nb.no/search (last visited: 2023/04/13)

[3] https://www.nb.no/ngram (last visited: 2023/04/13)

[4] https://hdl.handle.net/21.11146/76

2 NB DH-LAB

The NB DH-LAB infrastructure[5] consists of 1) a large corpus of digital texts and metadata in the form of a database (not directly accessible), 2) a REST API granting restricted access to this database from the WWW in the form of HTTP requests, 3) a Python client for simple interaction with the REST API, and 4) web applications built using the Python client. The infrastructure subscribes to the FAIR principles: the objects in the platform are easily findable, globally accessible and persistent, ensuring re-usability. The infrastructure is built around open, documented APIs and thus built for interoperability, offered through the Norwegian Language Bank at the National Library of Norway (a CLARIN C-centre).

2.1 Corpus

As of 2023, NB DH-LAB maintains a text corpus of approx. 160 billion running words (mostly in Norwegian), making it one of the largest corpora in the world (see counts in Table 1). For example, the German Reference Corpus currently has a size of 55 billion tokens[6], and for the closed Google Books corpus used for the Google Books Ngram Viewer we assume 468 billion words for English in 2012 (cf. Lin et al., 2012, p. 170). At the same time, Norwegian is a fairly small language with its approx. 5 million speakers, so the corpus is large both in absolute and relative terms.

Material type	Documents	Tokens
Books	600,000	34,000,000,000
Newspapers	4,000,000	110,000,000,000
Journals	100,000	14,000,000,000

Table 1: Corpus counts

The written text sources in the collections of the NLN are scanned (if not born-digital), OCR'ed and represented as ALTO XML, a common format for OCR'ed text.[7] When creating a text corpus from the digitized objects in the DH-LAB, we extract text from the ALTO XML, tokenize it using a Norwegian text tokenizer and store it in a database (see below).

Crucially, each digitized object in the NLN gets its own Uniform Resource Name (URN), a persistent, globally unique identifier. Furthermore, each derived text object in the DH-LAB (e.g. text extracted from ALTO XML) gets its own URN, making it possible to handle multiple text extractions of the same digitized object. Thus even if a better OCR version is available from the library, the older version is still accessible using its identifier.

2.2 Fulltext API

At the core of NB DH-LAB is a database server with full-text. For indexing 160 billion tokens we considered and tried several alternatives, such as Corpus Workbench (CWB), but due to the sheer size of our digital material and considering that our library corpora are largely unannotated, we landed on a simple, but very performant solution using SQLite (see also Evert, 2010). SQLite hardly needs any setup and can be easily moved around. SQLite does not provide any sharding by default, instead we use separate database files for each partition (e.g. 20,000 books per database). In this way, we can easily add new texts without having to rebuild all indices, which is a very costly operation for a database with millions or even billions of rows.

Each database partition, then, contains a CoNLL style table with one row per token together with its reference in the corpus. In Table 2, an example from the full-text table is provided. We store the internal identifier together with the token and its position in the text.[8]

[5] https://hdl.handle.net/21.11146/88

[6] https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (last accessed: 2023/08/30)

[7] Born-digital material is delivered to the library mostly in the form of PDFs and since extracting text from PDFs is not straight-forward, these are normally OCR'ed as well.

[8] The integer identifier used here is part of the URN for the text (URN:NBN:no-nb_dhlab_100007393), in this case an English version of Ibsen's *A doll's house*. Using the URN resolver, http://urn.nb.no/URN:NBN:no-nb_dhlab_100007393, users get basic metadata for each object.

identifier	token	sequence_nr	paragraph_nr	page_nr
100007393	Hide	652	39	9
100007393	the	653	39	9
100007393	Christmas	654	39	9
100007393	tree	655	39	9

Table 2: Example of full-text table

Furthermore, we create virtual full-text search tables using the sqlite FTS5 extension[9] for each document and paragraph, each represented as a tokenized string built from the full-text table above. This allows for very fast and complex full-text queries on document and paragraph level. We also generate tables with unigrams and bigrams for each text, so that frequency lists can be exported and aggregated on corpus-level.

Since SQLite does (explicitly) not provide any server functionality, we implemented the server logic in Python. We maintain a global mapping table containing references to the objects with some basic (Dublin Core-style) metadata and the database files (partitions) they are stored in and query the databases in parallel. The database server is accessible via a REST API (https://api.nb.no/dhlab), documented in a Swagger interface (OpenAPI 2.0). A corpus in terms of the API is simply a list of identifiers (URNs) that are used in the various endpoints.

2.3 Python client

Most users will neither have the knowledge nor the interest in using the REST API directly, therefore, the DH-LAB maintains a Python client for the DH-LAB REST API (https://pypi.org/project/dhlab/), with documentation (https://dhlab.readthedocs.io/en/stable/). Additionally, we have built an R package (https://github.com/NationalLibraryOfNorway/dhlabR). We also provide sample Jupyter notebooks showcasing the basic functionality of the Python package (currently only available in Norwegian).

The Python client is intended also for users with limited Python knowledge, whereas most of the heavy-lifting is done on server-level (REST API). The Python package gives easy access to features such as:

- corpus builder with library metadata and content words
- frequency lists
- concordance viewer with a link to the digital library
- collocation analysis

Below we provide examples of these four functions (using version 2.26.1 of the package).[10] The code in Listing 1 will first create a corpus consisting of a sample of five books in English, with Henrik Ibsen as author and which are classified as fiction (Dewey 800 series):

```
import dhlab as dh
ibsen_corpus = dh.Corpus(doctype="digibok", author="Henrik Ibsen",
↪  ddk="8%", lang="eng", order_by="random", limit=5)
```

Listing 1: Build a corpus of a corpus of books using library metadata

[9] https://www.sqlite.org/fts5.html (last visited: 2023/04/13)
[10] https://pypi.org/project/dhlab/2.26.1/ (last visited: 2023/08/30)

The resulting corpus object contains basic metadata of the included documents and their identifiers and some simple functions for e.g. getting counts, concordances and collocations. Getting frequency lists for each text is as simple as show in Listing 2:

```
ibsen_corpus.count().head(5)
```

Listing 2: Extract counts of the top five tokens in the corpus

	100012508	100013007	100013529	100022188	100020384
.	3537	18164	18364	11235	88
,	2159	15677	6189	5158	4524
the	13	981	3315	2783	3424
to	669	53	2415	2256	1551
I	639	4157	68	27	1457

Table 3: Frequency lists (top five) from all books in the corpus (the header row contains textual identifiers)

Table 3 contains frequency lists (top five) for all texts in the corpus together with their identifiers (in the header). Concordances are provided via the `conc` function. The concordance window is limited to 25 words at API level (here a window of 10 is chosen) and concordances are not allowed to span paragraphs. In this way, we can share small portions of textual data without challenging copyright law or making it possible to reconstruct complete texts. The concordance output also contains a clickable link to the IIIF viewer in *Nettbiblioteket*, allowing the user to see more context given access to the object (for reasons of space, only the URN itself is shown here). Listing 3 produces Table 4 below:

```
ibsen_corpus.conc(words="snow", window=10, limit=5).show()
```

Listing 3: Show concordances for the word *snow* in the Ibsen corpus within a window of ten words

link	concordance
URN:NBN:no-nb_digibok_2010090320019	In the **snow** , high up in the wilds of...
URN:NBN:no-nb_digibok_2010101820024	In the **snow** , high up in the toilds of...
URN:NBN:no-nb_digibok_2010090320019	**snow** . He will , you can be sure ....
URN:NBN:no-nb_digibok_2010090320019	... show you a church Built of ice and **snow** .
URN:NBN:no-nb_digibok_2010090820016	... soul ' s pure as **snow** ! Såiling far and wide...

Table 4: Concordances for the word *snow* in the Ibsen corpus

Finally, users can extract co-occurence statistics for a word and do collocation analysis with the `coll` function. The column *counts* refers to the co-occurence of the node word and its collocate in the corpus. The relevance score is a variant of pointwise mutual information (see Johnsen, 2021) where the reference corpus is simply a list of the 50,000 most frequent words in the complete NLN text corpus. Listing 4 leads to table 5:

```
ibsen_corpus.coll(words="Peer",
 ↪  reference=dh.totals(50000)).show().sort_values(by="relevance",
 ↪  ascending=False).head(10)
```

Listing 4: Show collocations for the corpus

collocate	counts	relevance
Gynt	265	10906.021666
Peer	34	985.894250
are	31	182.172700
and	156	118.531038
as	40	74.834610

Table 5: Colllocations

The Python client (and the REST API) allows for many further applications, such as on-the-fly POS tagging and Named-Entity Recognition of texts using SpaCY, topic modeling and dispersion plots. For these functions, we refer to the documentation.

2.4 Web applications

The final piece in the DH-LAB corpus infrastructure is what we call our App Cloud, intended for users with no programming knowledge and for demonstration purposes. These web-applications consume the REST API using the Python client and are built using streamlit.io, allowing rapid application development. Some of the web applications are general-purpose, others are tailored for a specific research project. Working with a corpus is as simple as dragging an Excel document with textual identifiers and metadata into the app.

3 Concluding remarks

In this paper, we have given a short outline of the digitization project at the National Library of Norway and the corpus infrastructure built around it. We believe that our offerings are quite unique for a national library in Europe. Users of the infrastructure have limited access to nearly the complete written heritage of published Norwegian documents. In the future, we will explore handwritten material, web archives, but also further media types such as imagery and audio data.

References

Birkenes, M. B., Johnsen, L. G., Lindstad, A. M., & Ostad, J. (2015). From digital library to n-grams: NB n-gram. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, 293–295. https://aclanthology.org/W15-1839

Evert, S. (2010). Google web 1T 5-grams made easy (but not for the computer). *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, 32–40. https://aclanthology.org/W10-1505

Johnsen, L. G. (2021). Term distance, frequency and collocations. In *Language and text: Data, models, information and applications* (pp. 21–36). Benjamins.

Lin, Y., Michel, J.-B., Aiden Lieberman, E., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books NGram corpus. *Proceedings of the ACL 2012 System Demonstrations*, *13*, 169–174. https://aclanthology.org/P12-3029

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

CORLI CLARIN K Centre: Development and Perspectives

Christophe Parisse
Modyco
University of Paris Nanterre, France
cparisse@parisnan-terre.fr

Céline Poudat
BCL
University Côte d'Azur, France
celine.poudat@univ-cotedazur.fr

Abstract

One of the primary objectives of the CORLI CLARIN K Centre is facilitate collaboration among researchers in the field of language sciences. The center aims to foster the development of projects that might be beyond the scope of individual researchers and to provide access to cutting-edge digital tools that enhance their scientific endeavors. These tasks are achieved by providing support and training in the utilization of modern digital tools designed for tasks such as corpus creation, annotation and data analysis. However, there are instances where the existing tools prove insufficient for the demands of language research. This can occur when these tools lack necessary functionalities, are unavailable in the required format or do not align with specific research needs. In light of these challenges, we will introduce two ongoing projects within CORLI that are focused on bridging the gap between researchers, technology and data:

- Open French Corpus: A centralized platform for accessing and utilizing existing corpora with shared tools.
- Collaborative annotation: use and improve existing tools; connect researchers, educators and students; develop a collaborative resource.

1 Introduction

The CORLI CLARIN K Centre[1] (Parisse et al., 2017; Soroli et al., 2020) was created in 2020. Comprised of members from over 20 French research labs and 15 Universities, the consortium is part of the lage French infrastructure Huma-Num. This infrastructure is dedicated to assisting researchers in the Humanities to use all types of digital data and tools. The CORLI CLARIN K Centre mainly aims to respond to users' needs regarding data and tools. We offer information, training and facilitate discussions and among academic users. These efforts aim to foster the development of projects and recommendations across various research areas involving language corpora.

It is during these panels and at our annual CORLI conference that proposals surfaced, addressing user needs in two domains of significant interest to the linguistic research community. In both instances, the projects CORLI undertakes build upon existing tools or standards that are already used and endorsed by the research community. However, these tools and standards fall short of fully meeting researchers' requirements. This gives rise to two main situations in which CORLI can play a pivotal role in assisting these situations to be resolved.

1) **Integration of tools and data**: Combining various tools or data components into a more comprehensive tool or dataset.

[1] https://www.clarin.eu/blog/tour-de-clarin-french-clarin-knowledge-centre-corli-corpora-languages-and-interaction

2) **Project expansion or adaptation**: Extending or modifying a research project to suit different scenarios or contexts.

2 An Open French Corpus

The initial proposal involved providing unified access to a high-quality corpus of the French language. Numerous independent corpora exist for French, often stemming from funded project. While these corpora might be substantial in size, the tools supporting their use might no longer be maintained once the project concludes. Additionally, issues with corpus format maintenance can hinder their usability. Alternately, some corpora result from collaborative efforts of researchers or laboratories over multiple years. While these corpora tend to exhibit exceptional quality, they might not attain the scale of those funded by larger projects. Nevertheless, they offer the advantage of being more recent, undergoing modifications and extensions. Diverse corpora are also present, such as those created for doctoral theses or cultivated by individual researchers through dedicated efforts.

In any case, these corpora are often clearly identified and accessible in well-known formats. Some are securely stored in institutional archives such as ORTOLANG or COCOON, university repositories, or even private repositories in some cases. In all cases, the corpora are accompanied by scientific publications describing their creation, format, and objectives.

CORLI's mission is to gather these scientifically validated sources into a single repository. Our objective is to standardize the format across all data and enable their utilization through a common set of tools. The efforts of CORLI are directed along three primary avenues:

1) **Harmonizing Metadata**: Establishing a fundamental metadata structure that can be universally applied for processing all data.
2) **Data Format Conversion**: Transforming data into both raw text and TEI (Text Encoding Initiative) formats, catering to distinct usage scenarios.
3) **Providing Processing Tools**: Furnishing tools capable of processing, querying, and displaying the complete dataset.

Whenever feasible, our approach entails automated conversion and processing. This approach not only facilitates the integration of future data—whether newly deposited into official repositories or updates to existing data—but also ensures efficiency in achieving our goals.

3 Collaborative annotation

Collaborative annotation plays a pivotal role within the linguistic community. The availability of extensive language corpora presents a challenge, as the task of editing and annotating an entire language dataset can be overwhelming for an individual. The process of corpus annotation demands significant time investment, ideally distributed across multiple contributors to make it feasible. While projects with substantial financial support can manage this, individual researchers, groups, or even entire laboratories face constraints.

Even in cases where annotation is performed automatically, there remains a need for manual oversight to verify and analyze the output of automatic processing. As a result, collaborative annotation has become a necessity in numerous instances.

Collaboration can take different forms. It may involve multiple skilled users, where the collaborative tools primarily facilitate data sharing and prevent redundant annotations. Alternatively, collaboration can include less experienced users. Here, the collaborative tools must enable researchers to compare several annotations and to resolve inconsistencies through an adjudication process.

The effectiveness of collaborative annotation is also influenced by the data format. While original data may be in text or TEI formats, tools already exist for editing such data, and CORLI will endeavor to leverage these existing resources. More intricate situations arise when the original data takes the form of images (such as handwritten materials or low-quality documents). In such cases, the image must be displayed, and annotations need to be linked to specific portions of the image. Lastly, annotations could also be generated automatically and subsequently checked manually.

The range of scenarios requires tailored annotation environments. Recognizing the impossibility of a one-size-fits-all solution, our approach is to focus on enhancing and utilizing three distinct tools that cater to the specific requirements of CORLI-affiliated laboratories:

1. TACT (Transcription and Annotation Collaborative Tool): The TACT initiative (https://tact.demarre-shs.fr/) centers around a web platform designed for transcribing and annotating text corpora. Our objective is to enrich the platform's capabilities, enabling it to facilitate crowdsourced transcription of parliamentary data.
2. INCEption (Integrated Corpus Exploration): INCEption (https://inception-project.github.io/) is a web-based annotation tool for corpus data. Our aim here is to improve the conversion of external data into the tool's internal format, encompassing data with pre-processed syntactic or semantic annotations. To this end, we've developed a converter (https://corli.huma-num.fr/convinception/#/sax2) that allows users to import and export XML corpora to and from Inception.
3. GUM (Grammatical Universal Dependencies Multilayer Corpus): The objectives of the CORLI-GUM are twofold. Firstly, we seek to foster collaboration among researchers, teachers, and Master's students. Researchers often lack annotators for their projects, while teachers express interest in engaging students in ongoing annotation initiatives. This approach enhances students' learning experiences and motivation. Secondly, we endeavor to develop an open-source, multi-layer corpus of richly annotated texts (https://gucorpling.org/gum/), following the example set by Amir Zeldes in 2017. The resulting resource will be made accessible to the wider community.

4 Conclusion

The ongoing objective of the CORLI CLARIN K Centre is to continue assisting researchers in France by providing access to the most effective linguistic tools, including those furnished by the CLARIN infrastructure. Simultaneously, we persist in the development, enhancement, and utilization of tools that researchers and laboratories within the CORLI network require for their work.

In alignment with our previous approach, we remain committed to augmenting existing tools rather than starting from scratch. This methodology not only proves more cost-effective but also resonates better with the community, as it aligns with their current practices and needs.

The CORLI CLARIN K Centre's operations are made possible through funding from the CORLI consortium, facilitated by a Huma-Num grant (https://www.huma-num.fr/les-consortiums-hn/), extending until 2024. This grant empowers us to continue our mission in supporting linguistic research and tool development within the community.

References

Christophe Parisse, Céline Poudat, Ciara Wigham, Michel Jacobson, Loïc Liégeois. CORLI: A Linguistic Consortium for Corpus, Language and Interaction. CLARIN Annual Conference 2017, Sep 2017, Budapest, Hungary. ⟨halshs-01636943⟩

Efstathia Soroli, Céline Poudat, Flora Badin, Antonio Balvet, Elisabeth Delais-Roussarie, et al.. CORLI: The French Knowledge-Centre. CLARIN Annual Conference 2020, Oct 2020, Barcelone (virtual), Spain. ⟨hal-03091629⟩

Zeldes Amir. The GUM corpus: creating multilayer resources in the classroom. Language Resources and Evaluation 51, 581-612, Springer.

Annex: Non-exhaustive list of corpora to be included in the OFC

- Written language corpora
  - Scientext https://scientext.hypotheses.org/corpus
  - Scienquest https://corpora.aiakide.net/

    - Archives parlementaires https://archives-parlementaires.persee.fr/
    - Consortium CAHIER https://cahier.hypotheses.org/
    - E-CALM https://www.ortolang.fr/market/corpora/e-calm
    - Corpus 14 https://www.univ-montp3.fr/corpus14/
    - Democrat https://hdl.handle.net/11403/democrat
- Spoken language corpora
    - CFPP2000 http://cfpp2000.univ-paris3.fr/search.html
    - ESLO http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus
    - CHILDES https://talkbank.org/DB/
    - CT3-ORTOLANG https://ct3xq.ortolang.fr/ct3xq/interro
    - PFC https://public.projet-pfc.net/transcription/
- Multiple-format corpora
    - CEFC-Orféo https://orfeo.ortolang.fr/, http://orfeo.grew.fr/
    - CoMeRe http://hdl.handle.net/11403/comere

The SSH Open Marketplace and CLARIN

Alexander König
CLARIN ERIC
alex@clarin.eu

Laure Barbot
DARIAH-EU
laure.barbot@dariah.eu

Cristina Grisot
University of Zurich
CLARIN-CH
cristina.grisot@uzh.ch

Michael Kurzmeier
University College Cork
MKurzmeier@ucc.ie

Edward J. Gray
DARIAH-EU
IR* Huma-Num
edward.gray@dariah.eu

Abstract

The SSH Open Marketplace is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities: tools, services, training materials, datasets, publications and workflows. This proposal presents how this service can provide insights into the use of tools, methods and standards in the Social Sciences and Humanities communities in general, and for the CLARIN community in particular. The paper also describes how the SSH Open Marketplace can increase serendipity in the discovery of new methods and standards, by interlinking the resources and describing workflows. Because contextualisation is provided between the items of the catalogue, it is easy to understand and assess the usefulness of a resource. Participants of the CLARIN Annual Conference 2023 are introduced to the functioning of the SSH Open Marketplace and are invited to contribute to the creation of new or enrichment of existing records.

1 Introduction

The Social Sciences and Humanities Open Marketplace (SSH Open Marketplace) - marketplace.sshopencloud.eu - is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities: **tools, services, training materials, datasets, publications and workflows**. The SSH Open Marketplace showcases solutions and research practices for every step of the research data life cycle. In doing so, it facilitates discoverability and findability of research services and products that are essential to enable sharing and re-use of workflows and methodologies.

The creation of the SSH Open Marketplace was funded by the Social Sciences and Humanities Open Cloud (SSHOC) project [1] which supported the integration and consolidation of thematic e-infrastructure platforms in preparation for connecting them to the European Open Science Cloud. (EOSC)[2]. The overall objective of the SSHOC project was to realise the Social Sciences and Humanities component of EOSC. As a domain-oriented discovery portal and the aggregator of the SSHOC project, the SSH Open Marketplace, contributes directly to the EOSC, supplementing existing services such as the EOSC Catalogue and Marketplace, and facilitating the fluid exchange of tools, services, data, and knowledge. As a continuation of the SSHOC project and to sustain its outputs, 5 ESFRI Landmarks CESSDA, CLARIN, DARIAH, ESS and SHARE have signed a Memorandum of Understanding for the establishment of the **SSH Open Cluster**. This cluster acts as an umbrella for the SSH Open Marketplace organisation and activities. More generally, the collaboration between the SSH Open Marketplace stakeholders (funders, providers, moderators or contributors) ensures that these cataloguing and contextualising efforts are meaningful, notably because they are undertaken by and serve humanities researchers.

The SSH Open Marketplace is one of the 33 Key Exploitable Results of the SSHOC project, and CLARIN, DARIAH and CESSDA decided to ensure the sustainability of the service after the end of the project. They act as a Governing Board for the SSH Open Marketplace and define the Marketplace

[1] See the SSHOC project description: https://cordis.europa.eu/project/id/823782

[2] See EOSC description on the EOSC Association website: https://eosc.eu/eosc-about

strategic policy with regards to scientific, technical and managerial matters. In that context, two institutions act as service providers on behalf of these ERICs: the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) of the Austrian Academy of Sciences providing hosting and maintenance for the service; and the Poznan Supercomputing and Networking Center (PSNC), affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences, providing the data ingestion pipeline as well as maintenance for the service. The SSH Open Marketplace can also count on an Editorial Board, composed of 17 members[3], to ensure the day-to-day maintenance and (meta)data quality. Liaising with service providers and the end-users of the service (SSH researchers and support staff for researchers), the Editorial Board ensures the technical running of operation, the effectiveness of the curation process and the editorial policy's successful implementation.[4].

In sum, CLARIN has been heavily involved in the SSHOC project and is a founding partner for the continuation of the project in the form of the SSH Open Cluster. The SSH Open Marketplace is one of the key elements of research empowerment and discovery with which CLARIN is concerned. The special focus on contextualistion of the resources in the Marketplace can act as a complementary discovery tool to CLARIN's Virtual Language Observatory (VLO) - which includes a much larger number of items, but presents a lot less context - and the CLARIN Resource Families - which contain a much smaller number of items, but therefore can be even more extensively curated and contextualised.

2 Presentation of the SSH Open Marketplace

2.1 Guiding principles

While planning and building the SSH Open Marketplace three main pillars were identified, and these remain essential for its ongoing operation and future development. These pillars are:

Curation - The service thrives on a curation process that makes it easy to discover the most appropriate and up-to-date results for each request, so that researchers can discover the best resources for the digital aspects of their work. The curation process relies on three components: automatic ingest and update of data sources; continuous curation of the information by the editorial team and – most important – contributions from users, the SSH research community.

Community – The content available in the SSH Open Marketplace and its contextualisation is the result of collaborative work that is characterised by a user-centric approach. Features that allow contributions are implemented to ensure that the portal mirrors real research practices.

Contextualisation – The portal puts all items into context: each solution suggested is linked to other related resources (e.g. a tutorial showing how to use a tool, a tool used in a workflow, a publication presenting research results produced using a given service). This contextualisation enhances the usefulness of the SSH Open Marketplace by showing how all these parts of the research process intertwine, and ensures users receive the maximum possible benefit from all its contents.

2.2 Inclusion criteria

In order to guide users who wish to add resources to the SSH Open Marketplace, inclusion criteria and related guiding questions are enforced:

The relevance of the resource. The question to ask is: *will this resource be relevant to the SSH scientific community?* Thus, to be selected, any resource must fulfil at least two criteria: (1) scientific relevance and usefulness for SSH research and researchers and (2) pertinence to the digital methodologies used within the SSH landscape.

The technical status of the resource. The question to ask is: *is the resource current, supported, and ideally open?* The SSH Open Marketplace favours the uptake of Open Science workflows and open research practices. Software resources are preferably built upon open source solutions. Nonetheless, given that the SSH Open Marketplace seeks to mirror actual research practices, commercial or non-current resources are also referenced where these are relevant for the scientific community.

[3] see this page on the Markeplace website: https://marketplace.sshopencloud.eu/about/team

[4] For a detailed version of the sustainability plan, the report on Marketplace governance (Petitfils et al., 2021) can be consulted

The degree of compliance with Open Science requirements of the resource. The question to ask is: *is the resource FAIR – Findable, Accessible, Interoperable and Re-usable - or contributing to the uptake of Open Science best practices?* The SSH Open Marketplace maximises the findability and re-use of data, and guides users towards tools, services or training materials that can help them in their FAIRification of workflows.

The uniqueness of the resource. The question to ask is: *is the resource already in the Marketplace?* If yes, there is no need to add it again, either as an individual item or with a source. Users are invited to enrich these existing items.

Thanks to these inclusion criteria, the quality, the FAIRness and the relevance of the resources added on the SSH Open Marketplace is guaranteed. This is an essential advantage for researchers who use the SSH Open Marketplace to discover resources which originate not only in their discipline but also from outside their own discipline. For example, a scholar who studies history and who uses digital methods to examine old documents, can easily discover on the SSH Open Marketplace the *Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools*. [5]

2.3 Item types

There are 5 main content types on the SSH Open Marketplace, which are considered to be representative for the large array of digital resources that can be found on this discovery platform.

Tools and services which refer to all sorts of digital materials and products, such as software, applications, programs, websites, programming libraries and APIs, that make tasks easier to execute. The trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files *UDPipe*[6] is an example of a tool provided by CLARIN.

Training materials are tutorials, lessons or didactic resources explaining how to perform an action or highlighting the potential learning outcomes gained from using that material. For example, the *CLARIN Hands-on Tutorial on Transcribing Interview Data* [7]focuses on the role of automatic speech recognition – what are the opportunities, what are the pitfalls and to where can it be applied successfully.

Workflows are sequences of steps that one can perform on research data during their lifecycle. Workflows can be achieved by using diverse tools, resources and methods, and useful resources are connected to each step. For example, *Intertextuality phenomena in European drama history* [8]is a workflow composed of 4 steps useful for analysing the relationships between the characters in a drama based on monologue/dialogue.

Datasets are defined as an organised collection of data. They are generally associated with a unique body of work, typically covering one topic at a time and are treated as a single unit by a computer. The SSH Open Markeptlace indexes a lot of CLARIN resources, for example the *DK-CLARIN Reference Corpus of General Danish* [9]

Publications are defined as research results published in academic journals or non-peer-reviewed publication repositories such as Zenodo. The SSH Open Marketplace references only publications that can be connected to other resources (i.e. tools and services, training materials, workflows or datasets). For example, you can find a paper on *Using TEI, CMDI and ISOcat in CLARIN-DK*[10] or the *Dublin Core Metadata Schemas* [11] on the SSH Open Marketplace.

[5]Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools. Version 1 Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/training-material/duVII1

[6]UDPipe. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/tool-or-service/F7K42P

[7]SSHOC Webinar: CLARIN Hands-on Tutorial on Transcribing Interview Data. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/training-material/ITNpCC

[8]Intertextuality phenomena in European drama history. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/workflow/DMJlzG

[9]DK-CLARIN Reference Corpus of General Danish. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/dataset/iZT6Ua

[10]Dorte Haltrup Hansen, Lene Offersgaard, Sussi Olsen (2022): Using TEI, CMDI and ISOcat in CLARIN-DK. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/publication/4jQvZ5

[11]DCMI Schemas. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/publication/6kYac0

2.4 Moderation and Curation

With a population of approx. 6000 items, aggregated from 15+ trusted sources, the SSH Open Marketplace relies on community curation - i.e. contributions from the research communities in SSH and from the Editorial Board - to ensure the catalogue entries remain up-to-date and useful for SSH researchers, the end-users of the portal. Furthermore, curation routines, mixing automatic and manual tasks, are set up to ensure and continuously improve (meta)data quality. Indeed, in order to gain an overview of the SSH Open Marketplace data and to perform some analysis to prioritise the curation tasks and improve the Marketplace data quality, a Python library and a set of Jupyter notebooks have been created[12]. These flexible scripts allow moderators and administrators to query the SSH Open Marketplace with advanced parameters and filters and, in some cases, to write back to the system to flag some items for curation in the editorial dashboard.

3 The SSH Open Marketplace and CLARIN

3.1 CLARIN resources within the SSH Open Marketplace

As has been said, the SSH Open Marketplace has been populated from a wide variety of sources. Of the 15+ original sources of the Marketplace, two come from the CLARIN world. the linguistic tools from the Language Resource Switchboard (LRS) (Zinn, 2018) and the tools, corpora and lexical resources collected in the CLARIN Resource Families (CRF) (Fišer et al., 2018). In both cases the original metadata has been mapped to the Marketplace Data Model (Ďurčo et al., 2021). As both the LRS and the CRF are very active, which means that items are constantly being added or updated (and in some cases also removed), the SSH Open Marketplace team has decided for a continuous ingest, i.e. to regularly re-harvest them to reflect changes at the source in the Marketplace. Recently, CLARIN has strengthened its focus on training, especially in the context of the UPSKILLS project[13], and as one of the outcomes training materials have been created or collected. To increase their discoverability, these training materials are being added to the SSH Open Marketplace, either manually or via the SSH Training Discovery Toolkit[14], which is also regularly harvested as a source by the Marketplace.

3.2 Future plans

First, we will look into improving the connection of the Marketplace with the VLO, which is a vast discovery portal including almost a million metadata records harvested from 47 CLARIN Centres and various non-CLARIN sources like Europeana or ELRA Indeed, the SSH Open Marketplace by its nature extends beyond the CLARIN world and it could be interesting to investigate in what way the SSH Open Marketplce could completement the VLO, the CLARIN Resource Families and the tools in the Language Resource Switchboard both from a technical point of view (i.e. mutual harvesting) as from the points of view of increasing the findability and accessibility of language data. Second, the SSH Open Marketplace extends CLARIN's quite complex infrastructure of discovery portals, which already includes the VLO, the Language Resource Switchboard and the CLARIN Resource Families. This multitude of discovery portals can be confusing for researchers or developers that would like to include information about their resource into the CLARIN infrastructure and want to ensure the maximum visibility for the community.

The same could be the case for those researchers who aim to discover resources. These users currently have at their disposal more and more discovery platforms useful to access resources as data, tools and services. To diminish the risks of confusion, it is therefore planned to create a guide that clearly outlines the various options of discovery portals, with their similarities and complementarities, to better inform and guide users who want to share their resources, as well as users who search resources. For instance, while the SSH Open Marketplace is similar to the CLARIN discovery platforms, with respect to discovering language data and tools, it presents the advantage of showcasing these resources in a contextualised manner. Last but not the least, in opposition to all other discovery engines, the SSH Open Marketplace

[12]This library and the set of notebooks has been created by Cesare Concordia (CNR-ISTI) and is available under: https://github.com/SSHOC/marketplace-curation

[13]see https://upskillsproject.eu/

[14]see https://training-toolkit.sshopencloud.eu/entities?search=clarin

proposes a unique type of resource, which becomes more and more important when it comes to reproducible research: workflows.

References

Ďurčo, M., Barbot, L., Illmayer, K., Karampatakis, S., Fischer, F., Moranville, Y., Ocansey, J. T., Probst, S., Kozak, M., Buddenbohm, S., & Yim, S.-B. (2021). 7.2 marketplace – implementation. https://doi.org/10.5281/zenodo.5749465

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's Key Resource Families. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*. https://aclanthology.org/L18-1210

Petitfils, C., Dumouchel, S., Larrousse, N., Gray, E. J., Barbot, L., Roi, A., Ďurčo, M., Illmayer, K., Buddenbohm, S., & Parkola, T. (2021). D7.5 marketplace - governance. https://doi.org/10.5281/zenodo.5608487

Zinn, C. (2018). The Language Resource Switchboard. *Computational Linguistics*, *44*, 1–13. https://doi.org/10.1162/coli_a_00329

CLARIN-IT: texts, documents and new contexts

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

Angelo Maria Del Grosso
ILC "A. Zampolli" CNR
Pisa, Italy
angelo.delgrosso@ilc.cnr.it

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

Francesca Frontinini
ILC "A. Zampolli" CNR
Pisa, Italy
francesca.frontini@ilc.cnr.it

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

Abstract

In recent years, CLARIN has increasingly broadened its interest from linguistic resources to textual resources relevant to digital humanists. This new and attractive scenario requires new technologies for texts, variants, and digital representations of primary sources, their contexts, and complex relationships. VeDPH in Venice, CNR-ILC-CoPhiLab, and ILC4CLARIN in Pisa collaborate on DH projects. Together, they are working on extracting text from manuscript page images, annotating historical graffiti on georeferenced images, and identifying text in digital images of paintings and sculptures.

1 Introduction

The acronym CLARIN is an acronym for Common Language Resources and Technology Infrastructure, and it reflects the fact that the principal community to which it addressed its activities was originally composed of linguists and computational linguists. However, Language resources, such as dictionaries or wordnets, e.g. "Ancient Greek WordNet" (Bizzoni et al., 2014)[1], and textual resources, such as literary or documentary corpora, e.g. "Cretan Institutional Inscriptions" (Vagionakis et al., 2022)[2] are complementary instruments to study the immaterial cultural assets of a civilization. The new definition of CLARIN as "the research infrastructure for language as social and cultural data[3]" is consistent with this new and more extended vision of language and its contexts.

In the last years, CLARIN made an effort to meet the requirements of the Digital Humanities and Museums-Libraries-Archives communities (Del Fante et al., 2022). For this reason, CLARIN has broadened its boundaries toward the digital representation of cultural artifacts, in particular towards the digital representation of text-bearing objects, such as papyri, inscriptions, and manuscripts.

Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) are the necessary links between the digital representation of primary sources (i.e. facsimile images) and the conveyed textual content. CLARIN-IT is exploring the most suitable open OCR and HTR tools and services to be integrated into its infrastructure. At this stage, *eScriptorium* (Kiessling et al., 2019)[4] seems to be the best trade-off between openness of the licenses and recognition accuracy. *eScriptorium* exploits the IIIF protocol to seamless import facsimiles provided by authoritative digital archives (such as e-Codices[5] or Ambrosiana[6]) and digital libraries (such as Gallica[7] or the Bodleian Digital Library[8]). CLARIN-IT can help the Italian community of digital humanists in three ways. First, CLARIN-IT can host the manifests created by scholars and provide a permanent identifier (PID) through a handler service. Second, it can

[1] To deepen into the Ancient Greek WordNet see http://hdl.handle.net/20.500.11752/ILC-56

[2] To deepen into the Cretan Institutional Inscriptions project see http://hdl.handle.net/20.500.11752/OPEN-548

[3] The new definition is claimed on CLARIN HomePage: https://www.clarin.eu/

[4] The eScriptorium git repository can be found at https://gitlab.com/scripta/escriptorium

[5] e-Codices digital archive can be found at the following web address https://www.e-codices.unifr.ch/it

[6] Ambrosiana digital archive can be found at the following web address https://www.ambrosiana.it

[7] Gallica digital library can be found at the following URL https://gallica.bnf.fr

[8] Bodleian Digital Library can be found at the following URL https://digital.bodleian.ox.ac.uk

create a special interest group to work on the integration of OCR or HTR data expressed in standard formats, such as ALTO[9], and metadata expressed through the IIIF manifests[10]. Third, the K-Centre devoted to Digital and Public Textual Scholarship (DiPtext-KC[11]), is planning workshops and seminars about the IIIF best practices.

2 Collaboration between VeDPH, CNR-ILC-CoPhiLab and ILC4CLARIN

In the Italian scenario, Computational Linguistics and Digital Humanities have a long story of entanglements and separations (Buzzetti, 2019), of methodological sharings and high specialization in knowledge subdomains (Montemagni, 2013). Among the others, we focus our attention on a specific case of collaboration. The Venice Centre for Digital and Public Humanities[12] (VeDPH) of the Department of Humanities at the Ca' Foscari University of Venice has been working in synergy with the Collaborative and Cooperative Philology Lab[13] (CoPhiLab) of the CNR-Institute of Computational Linguistics "A. Zampolli"[14] (CNR-ILC) and with the B-Centre ILC4CLARIN[15] since its founding in 2019 (Fischer et al., 2023).

3 HTRoman and HTRogène (Italian Section)

VeDPH takes part to the projects HTRoman and HTRogène, lead by the University PSL (Paris Sciences et Lettres) and funded by Biblissima+[16]. The aim of the projects is the enlargement of the HTR-United[17] (Chagué et al., 2021) collection of accurate transcriptions of samples from Medieval manuscripts with heterogeneous layouts, written in different scripts for various Romance languages: ancient French, Occitan, Catalan, Castillan, Tuscan, and Venetian. VeDPH, supported by CNR-ILC, is working on the manuscripts produced in Italy. For HTRoman, a team of two proof-readers and a supervisor accessed eScriptorium[18]. The web platform integrates the following functionalities: a) image acquisition through uploading or through the IIIF protocol; b) layout analysis; c) text recogntion (through Kraken[19]); d) proof-reading; e) creation of a new model or of a fine-tuned model. Like the other national sections of the project, also the data related to the Italian section are available online under an open access license[20].

4 VeLa

Venezia Libro Aperto (VeLa, Venice Open Book) is a DH project lead by the Department of Humanities of Ca' Foscari University devoted to the digitization of the historical graffiti of Venice (De Rubeis, 2008), with the high priority of Ducal Palace. The project, currently funded by Biblissima+, involves VeDPH, MUVE[21], SABAP-VE-MET[22], CESCM[23] and CNR-ILC.

The project consists in the creation of a shared georeferenced database of all the graffiti of the Doge's Palace in Venice, from the 15th to the 20th century. Multiple transparent layers (according to different original hands and chronology) with the hand-drawn transcriptions are superimposed over the high resolution images of the graffiti.

Contextual metadata (related to place, shape, material, etc.) and textual data (related to transcriptions, named entities, etc.) will be encoded through VeLaDSL, a domain-specific language easily convertible in XML-TEI/EpiDoc to ensure the interoperability with other Biblissima+ projects.

[9] The ALTO XML document format is described by the following specifications https://www.loc.gov/standards/alto/
[10] The last API specifications are described at the following URL https://iiif.io/api/presentation/3.0/
[11] To deepen into the k-centre website see https://diptext-kc.clarin-it.it
[12] https://www.unive.it/pag/39287
[13] https://cophilab.ilc.cnr.it/
[14] http://www.ilc.cnr.it/
[15] https://ilc4clarin.ilc.cnr.it/
[16] https://biblissima.fr/
[17] https://htr-united.github.io/
[18] https://escriptorium.inria.fr/
[19] https://github.com/mittagessen/kraken
[20] https://github.com/HTRomance-Project/medieval-italian
[21] https://www.visitmuve.it/en/home/
[22] https://www.soprintendenzapdve.beniculturali.it/
[23] https://cescm.labo.univ-poitiers.fr/

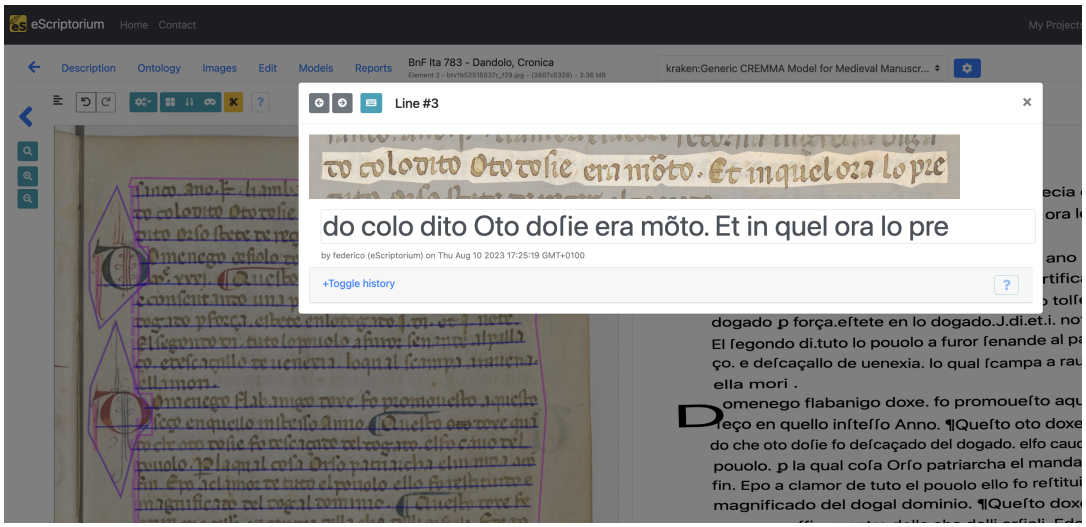


Figure 1: eScriptorium

5 Galleria Borghese

The project for the virtual museum of the Galleria Borghese in Rome (De Vincentis & Critelli, 2023) allows the navigation of highest resolution images, with a 360 degrees perspective, of the rooms of the gallery. Paintings and sculptures can be zoomed by exploiting the framework developed by the IIIF community. Images can be annotated according to the Web Annotation Data Model[24]. An interesting aspect of the project is the possibility to annotate regions of the digital images containing written texts (for example inscriptions on the basements of the sculptures or cartouches and scrolls within the paintings to transcribe the texts and to make them searchable and linkable to other textual sources. Part of data, navigation and annotation tools will be hosted in the new data center of the H2IOSC[25] consortium, located in Pisa.

6 On the possible integration in CLARIN-IT

The integration illustrated in Vagionakis et al., 2022 represents a model for other comparable DH projects. For the projects outlined in this contribution, we will follow the described strategy: a) the use of the repository of CLARIN-IT to describe both the data and the tool; b) the provisioning of the services through the CLARIN-IT servers, and c) a GitHub repository with data and software. This strategy requires to address some points and raises different questions.

In this section, we focus on (i) licenses for both data and tools; (ii) hardware and software requirements, and (iii) versioning.

Licenses (i) are of fundamental importance for a). As CLARIN, we can describe images and texts from the projects if they have been properly licensed. At this stage, licenses are not defined yet, but we may assume the images (at least a substantial subset of the entire collection) will be licensed under a CC-BY-SA[-NC]. This allows us to describe the (subset of) data in the ILC4CLARIN repository without defining a specifying license and specific access policies to the resources. The model described in Vagionakis et al., 2022 applies a total decoupling between data accessed by the CLARIN repository and data accessed by the application. We may follow the same strategy and limit the provisioning of the offered services to the hosting of such services, b). In such a case, ILC4CLARIN acts as the host of the IIIF servers but does

[24] https://www.w3.org/TR/annotation-model/
[25] https://www.h2iosc.cnr.it/

not interfere with the licences and access policies. The decoupling is important to dimension hardware and software as well, (ii). The ILC4CLARIN center will be dramatically improved during H2IOSC, both in terms of storage and GPUs. However, the ILC4CLARIN will be only a component of the H2IOSC project. b) allows us to host the IIIF services on different servers. Finally, (iii) and c) define a methodology and a workflow: we require developers to use GitHub as a repository for images and software. Every time a new release is available, a new version of the images in the ILC4CLARIN repository is submitted as well as a new version of the provided services. In this way, we guarantee the replicability: researchers can access previous version of data and software and replicate their experiments.

7 Conclusion

The collaboration of VeDPH in Venice with CNR-ILC-CoPhiLab and ILC4CLARIN in Pisa is an opportunity to work for the integration of language and textual technologies with image technologies in order to have a wider perspective on language as cultural data. Furthermore, the collaboration allows us to better address the following difficulties: a) to ensure the long term preservation and maintenance to projects based on the linkage of textual and visual resources and b) to constantly share the know-how among CLARIN, CNR and university, even when the projects receive small funding and consequently the turnover of human resources devoted to them is frequent.

References

Bizzoni, Y., Boschetti, F., Del Gratta, R., Diakoff, H., Monachini, M., & Crane, G. (2014). The making of Ancient Greek WordNet. *Proceedings of the 9th Annual Conference of LREC*.

Buzzetti, D. (2019). The Origins of Humanities Computing and the Digital Humanities Turn. *Humanist Studies & the Digital Age*, *6*(1), 32–58. https://doi.org/10.5399/uo/hsda.6.1.3

Chagué, A., Clérice, T., & Romary, L. (2021). Htr-united: Mutualisons la vérité de terrain! *DHNord2021-Publier, partager, réutiliser les données de la recherche: les data papers et leurs enjeux*.

De Rubeis, E., Flavia; Banterla. (2008). Scrivere sui pavimenti, scrivere sui muri: Materiali originali e riuso architettonico. In *Monasteri in europa occidentale (secoli viii - xi). topografia e strutture* (pp. 477–487). http://opac.regesta-imperii.de/id/1335049

De Vincentis, S., & Critelli, M. (2023). Mappare il museo in IIIF. Una combinazione di deep zoom e VR360 per la Galleria Borghese di Roma. *Proceedings of AIUCD2023*.

Del Fante, D., Frontini, F., Monachini, M., & Quochi, V. (2022). CLARIN-IT: An Overview on the Italian Clarin Consortium After Six Years of Activity.

Fischer, F., Boschetti, F., Del Grosso, A. M., Montefusco, A., Mancinelli, T., & Macchiarelli, A. (2023). Sinergie fra VeDPH e CNR-ILC in termini di condivisione della conoscenza e sostenibilità dei progetti digitali. *DH.22*. https://doi.org/10.48255/9788891328342.08

Kiessling, B., Tissot, R., Stokes, P., & Ezra, D. S. B. (2019). Escriptorium: An open source platform for historical document analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–19.

Montemagni, S. (2013). DH@ILC. In M. Agosti & F. Tomasi (Eds.), *Collaborative Research Practices and Shared Infrastructures for Humanities Computing. Proceedings of revised papers of the 2nd Annual Conference of the Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), Padova* (pp. 101–114). CLEUP.

Vagionakis, I., Del Gratta, R., Boschetti, F., Baroni, P., Del Grosso, A. M., Mancinelli, T., & Monachini, M. (2022). 'Cretan Institutional Inscriptions' Meets CLARIN-IT [ISBN: 978-91-7929-444-1 ISBN: 1650-3686]. In F. de Jong & M. Monachini (Eds.), *Selected Papers from the CLARIN Annual Conference 2021* (p. 189). CLARIN ERIC. https://doi.org/https://doi.org/10.3384/9789179294441

Documenting Corpus Annotation in CMDI: State of Affairs

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

Abstract

This paper discusses how the annotation of language corpora is documented in CMDI metadata. It shows that the most widely used CMDI profile for annotated corpora defines only one component related to annotation (i.e., for type), in contrast to less widely used profiles that define several components or elements (for tagset, tool, mode, segmentation level, and so forth). Furthermore, only a minority of corpora document annotation in the form of dedicated CMDI components; instead, such information is most often provided in an unstructured manner as part of the Free-Text Description.

1 Introduction

One of the aims of the CLARIN Resource Families[1] (CRF) initiative (Lenardič & Fišer, 2022b) is to contribute to the ongoing curation of language resources and tools hosted by CLARIN repositories. In this regard, CRF has periodically reviewed the provision of a few basic metadata categories, which are in the case of language corpora size, licence, and annotation. In a 2020 review (Lenardič & Fišer, 2020), it was for instance shown that while the resource size and licence were provided for 91–92% of the 558 corpora included in CRF at the time, information on annotation was included for a considerably smaller number – that is, only for 76% of them.

Crucially for these reviews, what has counted as being included is a mention of the relevant metadatum anywhere in the repository entry, be it as part of the Free-Text Description (FTD) accompanying each entry or anywhere else in the CMDI metadata. For this paper, what we now want to review is how annotation, which is one of the most important metadata fields for researchers seeking out language corpora, is documented in the form of CMDI components that are dedicated to annotation specifically.

CMDI,[2] which is CLARIN's solution for metadata modelling, addresses "the heterogeneous nature of the metadata landscape with a high degree of interoperability and reusability, both within and across communities" (Windhouwer & Goosen, 2022, p. 194). A metadata record based on CMDI consists of metadata components, themselves defined in terms of other constituents such as elements, attributes or other embedded components. CMDI profiles and their constituent components, which are stored in CLARIN's Component Registry,[3] can be reused and retooled whenever new resources are deposited in CLARIN repositories. CMDI also allows for semantic interoperability, as its "semantic layer can be used for harmonized processing and presentation of metadata records from many different sources" (Windhouwer & Goosen, 2022, pp. 198–199). Consequently, making use of CMDI components to document metadata such as annotation is important from the perspective of a distributed infrastructure such as CLARIN, both for the resource creators and users. Additionally, one of the requirements for a CLARIN centre to receive B-centre certification is that they offer CMDI (see Wittenburg et al., 2013, p. 7).[4]

[1] https://www.clarin.eu/resource-families

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

[3] https://catalog.clarin.eu/ds/ComponentRegistry/

[4] As noted by an anonymous reviewer, not all CLARIN centres have adopted CMDI yet (e.g., many but not all C-certified ones, see here: https://centres.clarin.eu/). It is worth noting though that there exist training materials for adopting CMDI (https://www.clarin.eu/content/component-metadata#training) as well as a CMDI Best Practice Guide (accessible via the previous URL); however, there still seems to be room for additional materials in the future, such as for written guidelines on the adoption of CMDI.

The paper is structured as follows. Section 2 presents the main CMDI profiles used for annotated corpora and the frequency at which the annotated corpora avail themselves of the components dedicated to annotation. Section 3 discusses the link between the use of annotation components and repository types. Section 4 concludes with a brief proposal.

2 Annotation CMDI Components in Practice

The following survey was carried out on 26 April 2023, when there were 696 corpora included in CRF and 431 (62%) of which provided information on annotation in some form or another in their repository entries. Table 1 lists the 9 most frequent CMDI profiles of the 431 annotated corpora.[5] By far the most frequently used profile is LINDAT_CLARIN, accounting for almost a third of all the annotated corpora. The third column lists the number of corpora whose metadata records employ at least one CMDI component dedicated to annotation. While most of the CMDI profiles define at least one such component, three profiles – namely, data, OLAC-DcmiTerms, and DGDCorpus – lack them altogether, hence the zero percentages of inclusion in the Table.

CMDI Profile	**Nr. entries**		**Incl. anno. cmp.**	
LINDAT_CLARIN	125	29%	1	1%
resourceInfo	64	15%	31	48%
media-corpus-profile	56	13%	56	100%
data	40	9%	0	0%
MDrecord_corpus	36	8%	25	69%
corpusProfile	26	6%	22	85%
OLAC-DcmiTerms	26	6%	0	0%
DGDCorpus	20	5%	0	0%
TextCorpusProfile	11	3%	10	91%
Other	27	6%	8	30%
Σ	431	100%	153	35%

Table 1: The most frequently used CMDI profiles for annotated corpora in CLARIN and the rates at which dedicated annotation components are employed

Overall, the metadata records of only a minority of the annotated corpora – that is, 153 (35%) out of the total 431 – provide information on annotation in the form of dedicated CMDI components. On the level of the individual profiles, what stands out is the LINDAT_CLARIN profile,[6] as there is only one corpus that makes use of the profile's sole component dedicated to annotation (i.e., **annotationInfo**) – this is the *DK-CLARIN Reference Corpus of General Danish* (Asmussen & Halskov, 2011), whose metadata record provides 4 values for the element **annotationType**; concretely:

- **annotationInfo**
  - **annotationType:** tokenization
  - **annotationType:** sentence and paragraph segmentation
  - **annotationType:** POS-tagging
  - **annotationType:** lemmatization

The rest of the 124 LINDAT_CLARIN corpora document annotation elsewhere, usually as part of FTD. For instance, the FTD of *Written corpus ccGigafida 1.0* (Logar et al., 2013) states that the "corpus is annotated with morphosyntactic descriptions (PoS-tagged) and lemmatised", further specifying that the "XML file has PoS (MSD) tags in Slovenian only, while the vertical file has tags both in Slovenian and English", while for *Corpus of the Contemporary Lithuanian Language* (Utka et al., 2017) it simply states that it is "morphologically annotated".

[5] The full overview is available here:
https://docs.google.com/spreadsheets/d/1l6If1H1GyI47wi4HviJ3Zgln0sTjSHXHHSaRoY1vDq8/edit?usp=sharing

[6] Conversely, all the corpora using media-corpus-profile avail themselves of the profile's annotation components; all of these are spoken corpora belonging to the same repository – i.e., the Bavarian Archive for Speech Signals; see the *Regional Variants of German - Junior* (BAS, 2017) corpus for one such example.

In contrast to LINDAT_CLARIN, the profiles resourceInfo, MDrecord_corpus, corpusProfile, and TextCorpusProfile define several components dedicated to annotation. To see this, consider the following excerpt from the metadata record of the part-of-speech tagged and lemmatised *PTPARL* corpus (Mendes, 2012), which uses the resourceInfo profile. The annotation is given as part of the **annotationInfo** component, which further embeds 3 subcomponents – **annotationManualStructured**, **annotationTool**, and **sizePerAnnotation** –, and is reused for both annotation levels separately. Part-of-speech tagging is thereby documented in terms of the segmentation level, tagset (for which a hyperlink is provided), mode (automatic instead of manual), tools (the MBT tagger and YamCha), and the size of the subset that is annotated with parts of speech. A reference is also given for the paper describing the annotation process.

- **annotationInfo** [component]
  - **annotationType:** morphosyntacticAnnotation-posTagging
  - **segmentationLevel:** word
  - **segmentationLevel:** wordGroup
  - **tagset:** http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_en.pdf
  - **tagsetLanguageId:** en
  - **tagsetLanguageName:** English
  - **annotationMode:** automatic
    - **annotationManualStructured** [component]
      - **documentInfo:** (Généreux et al., 2012)
  - **annotationTool** [component]
    - **targetResourceNameURI:** http://ilk.uvt.nl/mbt/
    - **targetResourceNameURI:** http://chasen.org/~taku/software/yamcha/
  - **sizePerAnnotation** [component]
    - **size:** 975,806
    - **unit:** tokens

Comprehensive documentation of annotation in this manner is typical of the 153 corpora whose records make use of dedicated annotation components. The other 278 corpora document annotation mostly as part of FTD; however, the issue with FTD is that the documentation is usually limited to type, such as part-of-speech tagging and lemmatization in the case of the aforementioned *ccGigafida* (Logar et al., 2013) corpus. Additional information, such as the tagset and annotation tools, often needs to be sought elsewhere outside the repository entry, typically in papers describing the corpora.

3 CMDI Profiles and CLARIN Repositories

Certain CMDI profiles are characteristic of specific repositories. LINDAT_CLARIN is used solely by the repositories that employ the CLARIN DSpace architecture (Straňák et al., 2020), which is a fork of the well-known DSpace system (Smith et al., 2003) with additional features implemented by LINDAT that makes it suitable for use as a data repository (e.g., a new licencing framework). This reworked variant of DSpace serves as the underlying architecture of for instance the repositories of LINDAT/CLARIAH-CZ,[7] CLARIN.SI,[8] CLARIN-IS,[9] and ILC4CLARIN.[10] Such DSpace-based repositories are actually the most common in CLARIN (Hajič et al., 2022), which is likely why LINDAT_CLARIN is the most frequently used profile. By contrast, MDrecord_corpus is used by the CLARIN:EL repository[11] while resourceInfo is for instance used by the FIN-CLARIN[12] and CLARIN Portugal[13] repositories. These latter repositories use or are based on the META-SHARE system (Gavriilidou et al., 2012) rather than DSpace.

[7] https://lindat.mff.cuni.cz/repository/
[8] https://www.clarin.si/repository/xmlui/
[9] https://repository.clarin.is/repository/
[10] https://dspace-clarin-it.ilc.cnr.it/
[11] https://inventory.clarin.gr/
[12] https://metashare.csc.fi
[13] https://portulanclarin.net/repository/search/

The main difference (for our purposes) between DSpace and META-SHARE is that the resource entries of the former do not provide a separate field to which annotation CMDI components could be mapped, in contrast to other types of metadata such as size, licence, authorship, resource type, and so on. This DSpace approach of skipping over annotation is architecturally reflected in the Virtual Language Observatory (VLO, Van Uytvanck et al., 2010),[14] where it is possible to search for resources by keyword, resource type, modality, etc., but not annotation type. For users, this presents a problem of recall especially if they are interested in the subprocesses of annotation such as mode (e.g., manual instead of automatic); one attested solution is to list such information as a keyword (see *Training corpus SUK 1.0*, Arhar Holdt et al., 2022, as an example), but this is not a principled approach. By contrast, the META-SHARE repositories present a complex field where annotation can be defined in a highly structured and recursive manner, listing subcomponents such as the tools, tagsets, and schemata for each annotation type (and even modality) separately.

4 Proposal

Annotation should occupy a more prominent role in DSpace repositories, and in the CLARIN infrastructure more generally. Ideally, it should be presented as part of its own field in all corpus repository entries rather than simply being consigned to FTD. We have seen that the LINDAT_CLARIN profile, which is used by CLARIN DSpace corpora, currently only defines 1 component for annotation – that is, **annotationInfo** – along with the embedded **annotationType** element, which is used for values such as *PoS-tagged* or *Lemmatised*. But as pointed out by an anonymous reviewer, it is unclear how valuable the annotation type is in and of itself if the repository entry does not provide additional metadata on information like the tagset used or the accuracy of the annotation, or at least an explicit reference to a place where such additional information is available.

We therefore suggest that the LINDAT_CLARIN profile be revamped by defining such additional components relevant to the otherwise quite complex annotation process, which would bring it in line with e.g. resourceProfile or MDrecord_corpus. DSpace repositories should also consider implementing a way to input such (potentially optional) metadata on annotation during the depositing process itself. Note that the rate of provision of CMDI metadata is also currently evaluated by the CLARIN Curation Module,[15] which tracks the coverage of certain CMDI components for separate collections (which often correspond to the entire set of resources offered by a repository) harvested by the VLO.[16] Concomitantly, the VLO team should strongly consider integrating a facet dedicated to annotation, so that at least annotation type, which would be mapped from the relevant CMDI components, can be queried with VLO's faceted search.

But in lieu of such a solution, which would entail a partial restructuring of the existing repositories and thus might not be straightforwardly actionable in an infrastructure-wide scope, we also suggest that repositories adopt a consistent set of recommendations for describing annotation as part of FTD. Such documentation should be done in as comprehensive a manner as possible, especially if the metadata records make use of a profile with few relevant components (i.e., LINDAT_PROFILE) or none at all (i.e., data, OLAC-DemiTerms, DGDCorpus). What should be documented in FTD is not only the basic types, as is the practice of the majority of existing cases, but also all the subcomponents relevant for the end user, such as mode, tagset, theoretical framework (e.g., LFG vs. Universal Dependencies for syntactic annotation), and annotation tool.

In Lenardič and Fišer (2022a),[17] we present precisely such a set of qualitative guidelines that zone in on the identified gaps in the existing provision of annotation (and other) metadata, both in the case of those records that do avail themselves of bespoke CMDI components and those that resort to FTD. We believe that a cross-centre adoption of such recommendations is important because post-hoc curation of published resources requires a significantly higher investment of effort and time by the repository administrators. However, the operative word here is *recommendations* rather than necessary requirements

[14] https://vlo.clarin.eu/
[15] https://curation.clarin.eu/
[16] https://curation.clarin.eu/collection
[17] See also https://office.clarin.eu/v/CE-2022-2138-qualitative-depositing-recommendations.pdf.

for a deposit, as too many required metadata fields in the submission form could also discourage potential depositors. In addition to guiding the depositors during the submission process, such recommendations can also be of help to repository administrators, who need to evaluate the submission-in-progress from the perspective of the provided metadata as well as its qualitative documentation.

Acknowledgements

Work on this paper has been supported by the Slovenian Research Agency research programme *P6-0436: Digital Humanities: resources, tools and methods* (2022-2027) as well as the CLARIN Resource Families Initiative.

References

Arhar Holdt, Š., et al. (2022). *Training corpus SUK 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1747

Asmussen, J., & Halskov, J. (2011). *DK-CLARIN reference corpus of general danish*. CLARIN-DK-UCPH Centre Repository. http://hdl.handle.net/20.500.12115/36

BAS. (2017). *Bas regional variants of german – juveniles*. Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität München. http://hdl.handle.net/11022/1009-0000-0004-AE1D-9

Gavriilidou, M., et al. (2012). The meta-share metadata schema for the description of language resources. *Proceedings of LREC 2012*, 1090–1097.

Généreux, M., et al. (2012). A large portuguese corpus on-line: Cleaning and preprocessing. *Proceedings of PROPOR 2012*, 113–120. https://doi.org/10.1007/978-3-642-28885-2_13

Hajič, J., et al. (2022). Lindat/clariah-cz: Where we are and where we go. In *Clarin: The infrastructure for language resources* (pp. 61–82). https://doi.org/10.1515/9783110767377-003

Lenardič, J., & Fišer, D. (2020). Extending the clarin resource and tool families. *Proceedings of the CLARIN Annual Conference*, 1–5. https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf

Lenardič, J., & Fišer, D. (2022a). Clarin depositing guidelines: State of affairs and proposals for improvement. *CLARIN Annual Conference Proceedings*, 48–52. https://office.clarin.eu/v/CE-2022-2118-CLARIN2022_ConferenceProceedings.pdf#page=54

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

Logar, N., et al. (2013). *Written corpus ccGigafida 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1035

Mendes, A. (2012). *Ptparl corpus*. PORTULAN Clarin. https://hdl.handle.net/21.11129/0000-000B-D33C-4

Smith, M., et al. (2003). Dspace: An open source dynamic digital repository. http://hdl.handle.net/1721.1/29465

Straňák, P., Košarko, O., & Mišutka, J. (2020). Clarin-dspace repository at lindat/clarin. *Grey Journal (TGJ)*, *16*, 52–61.

Utka, A., et al. (2017). *Corpus of the contemporary lithuanian language*. CLARIN-LT digital library in the Republic of Lithuania. http://hdl.handle.net/20.500.11821/16

Van Uytvanck, D., et al. (2010). Virtual language observatory: The portal to the language resources and technology universe. *Proceedings of LREC 2010*, 900–903.

Windhouwer, M., & Goosen, T. (2022). Component metadata infrastructure. In *Clarin: The infrastructure for language resources* (pp. 191–222). https://doi.org/10.1515/9783110767377-008

Wittenburg, P., Van Uytvanck, D., Zastrow, T., Straňák, P., Broeder, D., Schiel, F., Boehlke, V., Reichel, U., & Offersgaard, L. (2013). *Checklist for clarin b centres, version 7.4*. http://hdl.handle.net/11372/DOC-78

Do Chatbots Dream of Copyright? Copyright in AI-generated Language Data

Paweł Kamocki
IDS Mannheim
Germany
kamocki@ids-mannheim.de

Toby Bond
Bird & Bird
London, UK
toby.bond@twobirds.com

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

Thomas Margoni
KU Leuven
Belgium
thomas.margoni@kuleuven.be

Abstract

For language scientists, a *prima facie* advantage of AI-generated data over human-created content is that AI outputs are generally regarded as free from copyright. This submission addresses this issue in some detail.

1 Introduction

2023 is the year of a rabbit according to the lunar calendar, but in Europe it is most likely to be remembered as the year of Artificial Intelligence. It is safe to say that such events as the launch of Chat-GPT (in November 2022) or of GPT-4 have already revolutionized the way in which language data are generated. This revolution has not been unnoticed by the CLARIN community. The new perspective that AI opens up, is to create fully synthetic data according to the specifications of a researcher.

In branches of science where data for language modelling is scarce, or access to it is limited by (usually copyright or data protection) laws, protected, e.g., medical sciences, behavioral sciences, etc., the researchers can ask an AI model to generate new data for large categories, thereby avoiding the legal barriers. The model can also be used for creating more data for small categories to make the data more balanced and less biased. However, the bias reduction needs to be verified so that the additional data does more than just amplify the prejudice or bias in the original data.

For language scientists, a *prima facie* advantage of AI-generated data over human-created content is that, as it is generally agreed upon, AI outputs are not protected by copyright. This abstract addresses this issue in some detail.

The main reason for the absence of copyright in AI-generated data is their lack of human authorship (Section 2). However, the re-use of certain AI outputs may be in a legal grey area (Section 3). The introduction of a property right in AI outputs is seen by some as an answer to the challenges presented by the development of generative AI (Section 4); however, little evidence of this is found in the UK, where computer-generated works have been protected by a property right since 1988 (Section 5).

2 Lack of human authorship as an obstacle to copyright protection of AI outputs

The argument commonly used to refuse copyright protection of AI-generated content is lack of human authorship. Indeed, the use of the word ‘author’ (Cambridge Dictionary: ‘a person who begins or creates something’) seems to indicate that only works created by humans can be protected by copyright.

Although human authorship is not expressly required by the Berne Convention, this landmark international treaty uses words like ‘nationality’ (esp. Art. 3), ‘honour’ (Art. 6bis(1)) and ‘death’ (Art. 6bis(2), 7, 7bis) to refer to the author, which clearly points at a human being. The same conclusion can be drawn from the EU Directive 2006/116/EC on Copyright Term (see esp. Art. 1(4)).

It is true that many copyright systems accept 'corporate ownership' of copyright, where copyright is held *ab initio* by a legal person and not the human author. This is for example the case under the work for hire doctrine, where copyright in a work created by an employee belongs *ex lege* to the employer.

Even if one overlooks the absence of human authorship, AI-generated outputs could not be protected by copyright, as they would not satisfy the originality criterion, at least in the European Union. According to the CJEU, a work is original if it constitutes its author's own intellectual creation, i.e. it reflects the author's personality, which is the case when the author can express their creative abilities by making free and creative choices[1]. Conversely, when technical considerations, rules and constraints leave no room for free choices, there can be no originality and therefore no copyright protection[2]. Since, arguably, AI outputs do not reflect the personality of their authors, and their generation follows technical constraints, they also do not meet the originality criterion and are not eligible for copyright protection.

The US Copyright Office followed this approach already in 2018 when it refused to register a machine-generated image[3]. In practice, however, the legal status of AI-generated works is more complex than it may *appear prima facie*.

3 Grey areas related to AI outputs

AI does not (yet) generate outputs autonomously; the generative process is always initiated by a human who prompts the application with an idea in their mind. At least according to the dictionary definition, this human initiator can still be referred to as 'author' ('a person who *begins* or creates something'), even though the actual expression of the work (protectable by copyright, unlike the initial idea) is generated (or at least assisted) by AI.

Drawing a line between outputs with sufficient human involvement to 'deserve' copyright protection ('AI-assisted') and those without it ('AI-generated') is an extremely delicate task (cf. the 4-step test in Hugenholtz and Quintais, 2021), and courts' views on this issue are susceptible of evolving over time.

Such was the case with, e.g., photography, which was admitted in the realm of copyright several decades after the technology was popularized, and even today it is not recognized in the Berne convention as equal with other types of works (Art. 7(4) allows for a shorter term of protection for photographic works). In early decisions involving photographs[4] courts emphasized the role of the human photographer in, e.g., selecting the lighting, a task that is (or at least can be) fully automated in modern digital cameras, which does not seem to affect copyrightability of digital photographs (Margoni, 2014). AI outputs may follow the same trajectory, and the degree of human involvement required by courts for copyright protection may be gradually lowered. After all, since the beginning of time, almost all forms of human expression have employed some form of technology, be it very rudimentary.

In its recent policy statement, the US Copyright Office (2023) also opted for a somewhat nuanced approach to registering AI-generated works. In the Office's view, merely prompting a machine is not enough to claim authorship in the output (no matter how elaborated the prompt, according to the Office it only functions as an instruction to a commissioned artist). However, copyright can be claimed where AI outputs are arranged by a human in a creative manner, or modified to a degree that meets the threshold of creativity. This is illustrated by the Office's decision regarding a comic book "Zarya of the Dawn"[5], in which all images were generated by AI. The comic book as such (the plot, the texts) were deemed eligible for registration, although individual AI-generated images were excluded therefrom.

Another grey area regarding copyright in AI outputs is linked to their relationship with the input data used to train the underlying model. Although the use of copyright-protected content to train AI models is generally (under certain conditions) allowed under the exceptions for Text and Data Mining (Kamocki et al., 2018), the legal status of AI outputs is rather unclear. Carlini et al. (2021) have shown that certain text data AI models may sometimes 'regurgitate' portions of training material, which contributes to significant lack of legal certainty regarding copyright status of such outputs, especially considering that according to the CJEU excerpts as short as 11 consecutive words may be protected by copyright in certain circumstances. Even without regurgitating verbatim copies of training data, some (e.g., Gervais,

[1] See esp. CJEU *Infopaq* (C-5/08) and *Painer* (C-145/10)
[2] CJEU *Football Dataco* (C-604/10)
[3] https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf
[4] See esp. *Burrow-Giles Lithographic Co. v. Sarony*, 111 U.S. 53 (1884)
[5] https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

2022) have argued that AI outputs are derivatives, derived from the training material, which would also impact their copyright status. This lack of legal certainty is illustrated by a recent US lawsuit, in which Getty Images sued Stability AI for allegedly using their images to train an AI model[6].

4 Towards (Property) Rights in AI Outputs?

In February 2023 it was reported that ChatGPT is listed as author or co-author of over 200 books available on Amazon (Nolan, 2023). One can only imagine the number of books and other texts that were 'secretly' generated by AI and passed as human creations. As purely AI-generated texts are generally in the public domain, they can fall victim to 'copyfraud', i.e. a false copyright claim (e.g. by simply signing an AI-generated text with one's name, as a pretended human author). In the current state of law, 'copyfraud', although certainly unethical, is usually not a punishable violation. In fact, the Berne Convention (Article 15(1)) and the EU Directive 2004/48/EC on the enforcement of IP rights (Article 5) establish a presumption of ownership for those whose name 'appear on the work in the usual manner'.

Already in the 1960s it was argued (Demsetz, 1967) that technological progress will necessarily be accompanied by the creation of new property rights, mostly to guarantee legal certainty of transactions and to prevent market failure. Indeed, in the last decades new property rights have been created, such as the *sui generis* database right, or the right in computer-generated works in the UK (see below).

Already in 2020 the European Parliament took the view that AI-outputs 'must' be protected under Intellectual Property Rights in order to encourage investment and improve legal certainty, and called the Commission to reform EU law accordingly. Such statements from the Parliament should, however, be regarded as devoid of any legal meaning. However, in a recent response[7], the Commission stated that 'the issue of AI-generated works does not deserve a specific legislative intervention'. Moreover, many European IP scholars criticize the idea of introducing new property rights (Bulayenko et. al, 2022).

On the other hand, in recent years the Commission was active in proposing governance-based (as opposed to property-based) regimes for data, including AI-generated data. This follows an attempt to introduce a data producers' right (Gangjee, 2022). These regimes, introduced e.g. by the Data Governance Act or the Data Act, are focused on rights of users, enabling access and portability of data (that companies want to keep 'secret'), rather than on recognizing monopolies (property rights) in the data (Margoni & Kretschmer, 2022). This can be a novel approach to regulating AI, both at the input end (e.g., by recognizing 'artist data', distinct from copyright in literary, artistic and scientific works), and at the output end.

For now, the re-use of AI outputs is mostly regulated by contracts, especially Terms and Conditions of related online services, which tend to vary significantly. For example, Terms of Use of ChatGPT allow for the generated content to be reused for any purposes, including commercial ones ('such as sale or publication'), with an important exception: the use of ChatGPT outputs to develop models that compete with OpenAI is prohibited. A similar prohibition can be found in Bard's Terms of Service. Bing's Terms of Use for its consumer-focused product only allow for the generated content to be reused 'for personal and non-commercial purposes'.

It should be noted here that if the outputs of these applications are not protected by copyright, copyright exceptions, including the TDM exceptions, cannot apply to them, and so the above mentioned Terms and Conditions cannot be overridden by such exceptions, as long as the contracts are enforceable.

Some language models, such as BERT or GPT-2, are also available under open source licences (Apache 2.0 and MIT, respectively), which impose no restrictions on the use of their outputs. However, more recent versions of GPT, starting from GPT-3, are publicly available only through a web API (i.e., subject to Terms and Conditions), and this trend is likely to continue with subsequent iterations of the most performant language models.

5 UK's Experience with Protection of Computer-generated Works

UK's Copyright, Designs and Patents Act of 1988 contains (since its adoption) a provision on computer-generated works (s9(3)). These works, defined as works 'generated by computer in circumstances such that there is no human author of the work', are protected by copyright (which, in the continental tradition,

[6] Getty Images (US), Inc. *v.* Stability AI, Inc. (1:23-cv-00135).

[7] https://www.europarl.europa.eu/doceo/document/E-9-2023-000479-ASW_EN.pdf (last access: 27.04.2023).

would be classified as a 'related' or 'neighbouring' right rather than copyright *stricto sensu*) for 50 years following their creation (s12(7)). The right belongs to 'the person by whom the arrangements necessary for the creation of the work are undertaken' (referred to as 'author'). Somewhat paradoxically, in order to qualify for protection, computer-generated works, like all other works, have to meet the criterion of originality (which historically was understood in the UK as involving a degree of 'labour, skill and judgement', but under the influence of the CJEU, a more author-centric approach to originality, presented above, was adopted). Similar provisions exist also in Ireland, New Zealand and South Africa.

Although it seems tempting to use this provision, adopted with the intention to regulate re-use of works such as satellite photographs, to AI-generated content, this has never been done by UK courts. In fact, case law involving this provision is extremely scarce, and the provision has been described as 'unclear and contradictory'. In a recent public consultation, the UK Intellectual Property Office listed computer-generated works as one of the issues to be addressed by the legislator. In its 2022 response, however, the government stated that, as there is no evidence that the provision is harmful, and 'any changes could have unintended consequences', especially given that the development of AI is still in its early stages. In the same statement, the government also declared that they will keep the provision under review and may remove, replace or amend it if the evidence supports this[8].

References

Bulayenko, O., Quintais, P. J., Gervais, D. & Poort, J. (2022). *AI Music Outputs: Challenges to the Copyright Legal Framework*. ReCreating Europe Report. https://doi.org/10.5281/zenodo.6405796

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A. & Raffel, C. (2021). Extracting Training Data from Large Language Models. *arXiv: 2012.07805.* https://doi.org/10.48550/arXiv.2012.07805

Demsetz, H. (1967). Toward a Theory of Property Rights. *The American Economic Review*, 57, 2, 347-359.

European Parliament. (2020). *Resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies* (2020/2015(INI))

Gervais, D. J. (2022). AI Derivatives: the Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines. *Seton Hall Law Review*, 53, 1111-1136. http://dx.doi.org/10.2139/ssrn.4022665.

Gangjee, D. S. (2022). The Data Producer's Right: An Instructive Obituary. [in:] Lim, E. & Morgan, P. (eds), *The Cambridge Handbook of Private Law and Artificial Intelligence*, Cambridge University Press.

Hugenholtz, P.B., & Quintais, J.P. (2021). Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? *International Review of Intellectual Property and Competition Law*, 52, 1190–1216. https://doi.org/10.1007/s40319-021-01115-0

Kamocki, P., Ketzan, E., Wildgans, J. & Witt, A. (2018). New exceptions for Text and Data Mining and their possible impact on the CLARIN infrastructure. *Selected papers from the CLARIN Annual Conference 2018*

Margoni, T. (2014). The Digitisation of Cultural Heritage: Originality, Derivative Works and (Non) Original Photographs (December 3, 2014). http://dx.doi.org/10.2139/ssrn.2573104

Margoni, T. & Kretschmer, M. (2022). A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology. *GRUR International*, 71(8), 685–701. http://dx.doi.org/10.2139/ssrn.3886695

Nolan, B. (2023). More than 200 books in Amazon's bookstore have ChatGPT listed as an author or coauthor. *Business Insider*, February 23, 2023. https://www.businessinsider.com/chatgpt-ai-write-author-200-books-amazon-2023

US Copyright Office. (2023). *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*. 16190 Federal Register, vol. 88, no. 51, 37 CFR Part 202.

[8] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents (last access: 27.04.2023).

Between Lexicon and Grammar: Towards Integrated Valencies for Bulgarian

Petya Osenova **Kiril Simov**
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Bulgaria
{petya, kivs}@bultreebank.org

Abstract

In this paper we share our experience with the assignment of semantic roles to the valency frames of Bulgarian verbs in the Bulgarian Valency Dictionary. Two types of resources are used: a) in-house ones like BTB-Wordnet for providing the lemma senses, and BulTreeBank as initial syntactically annotated corpus for valency frames extraction; b) external ones like VerbNet for English where steps of adaptation and localization of the semantic role set to Bulgarian were performed for each verb meaning. We also briefly outline our idea to cross-validate the lexicographic classes of verbs taken from BTB-Wordnet against the semantic roles assigned to the verb arguments in their specific contexts.

1 Introduction

In our latest work we consider creating integrated language resources like various types of dictionaries and corpora. This step is a key prerequisite for providing more complex language technologies and developing more complex linguistic research in multilevel and multilanguage directions. The main benefit from such an integrating (although not trivial) effort is the verification of distinct resources, simultaneous usage of complementary knowledge, and transfer of knowledge between resources within the same language or among different languages. Here we report on our first attempts at integrating a treebank (providing the syntactic structure through the initially extracted valency frames), a valency dictionary (providing the syntagmatic potential with verb meanings and related semantic roles) and a Wordnet of Bulgarian (providing the paradigmatic knowledge with lexical senses in an hierarchical manner). Valency dictionaries can be viewed as mini-grammars that connect lexica's potential with full-fledged grammars. For that reason they are a very valuable resource for a language. During the years many such dictionaries for various languages and in multilingual settings have been compiled with respect to differing linguistic approaches. Here we mention only some of the existing best practices. The interested reader can consult information about Croatian (Birtić et al., 2017), about Czech (Straňáková-Lopatková and Žabokrtský, 2002), about Polish (Przepiórkowski et al., 2014), about a multilingual setting (Di Fabio et al., 2019). At the time, a valency lexicon for Bulgarian was initially extracted from the original constituency version of BulTreeBank (Simov et al., 2004). This initial version of the resource followed the Valency Principle in Head-driven Phrase Structure Grammar (HPSG). This principle states that 'Unless the rules says otherwise, the mother's value for the VAL(ency) features (SPR(specifier), COMPS (complements), and MOD(ifier)) are identical to those of the head daughter' (Sag et al., 2003). The work on the Bulgarian Valency Dictionary had been first reported in (Osenova et al., 2012) and consequently it was developed further in some other works where the lexicographic classes from WordNet were also taken into account as semantic anchors for handling valencies. In those works however the semantic roles in valency frames were viewed in a general way, since only the lexicographic classes were used with their very prototypical roles like Agent, Patient, Experiencer, Theme, etc. The basic XML version of the resource has been submitted to ELEXIS and ELRA[1] repositories. However, the mentioned version 1.0 respects only

[1] https://catalogue.elra.info/en-us/repository/browse/ELRA-L0132/

the syntactic constraints of valency and only partially includes semantic constraints and semantic information. In this paper we would like to discuss an integrated annotation of verb valencies in Bulgarian with the usage of Pustejovsky's ideas on argument structure where he introduces three types of arguments: true (always realized syntactically), default (optional) and shadow (arguments being part of the lexical meaning of the verb) – (Pustejovsky, 1991); WordNet lexicographic classes[2]; verb senses from BTB-Wordnet and valency frames from VerbNet[3]. Thus the verb with its valencies is now linked to the respective semantic category and synset in BTB-Wordnet. The respective semantic roles are assigned to its arguments according to VerbNet. The frames follow Pustejovsky's ideas on argument structure such that the valency set includes all arguments depending on the lexical semantics and the selective potential of the verb. Our work adheres to the following definition of argument structure, namely 'As standartly assumed in generative frameworks, the argument structure specifies the number and type (both semantic and syntactic) of the arguments to a predicate. In GL[4], the AS[5] is seen as a minimal specification of a word's lexical semantics'(Pustejovsky and Batiukova, 2019, p. 81). In the same book the following example is given: *Mary built a house (from bricks)*. The arguments 'Mary' and 'house' are true, because they are obligatory realized in the text, while the argument 'from bricks' is default, because it is optional on the surface (p. 81). The shadow arguments are presupposed in examples like 'She buttered her toast with *butter'(p. 219) where the object duplicates the verb in its meaning and for that reason it is usually omitted unless further specified ('She buttered her toast with organic butter').

The knowledge transfer from English to Bulgarian has been performed in two ways: through the mappings between lexica in wordnets (BTB-Wordnet (Osenova and Simov, 2018) and Open English Wordnet (McCrae et al., 2019)) as well as through incremental localizations from VerbNet. The resulted valency resource is planned to be released as a stand-alone application and also as an integrated application in the platform 'All about words'[6].

2 The approach

Based on previous research – for instance (Osenova, 2022), we got evidence that Bulgarian prefers default arguments in comparison to true arguments and arguments in shadow. Thus, our representation of the Valency frames equals the HPSG representation of the argument structure (ARG-ST) which covers all the possible argument realizations in texts.

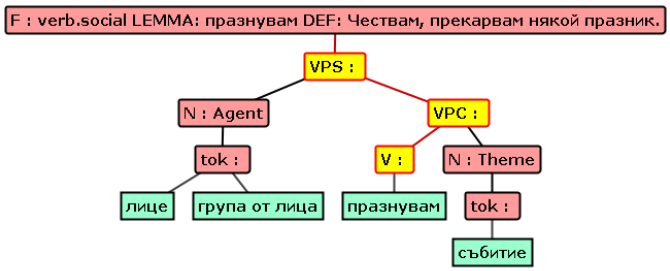


Figure 1: Valency of the verb 'celebrate'.

The lexicographic classes were transferred to the verbs in the respective meanings through BTB-Wordnet. The usage of the VerbNet resource for English, however, required localization and adaptation to the Bulgarian verbs. The fact that VerbNet has been mapped to a large extent to Princeton Wordnet and

[2]https://wordnet.princeton.edu/documentation/lexnames5wn
[3]https://verbs.colorado.edu/verbnet/
[4]GL = Generative Lexicon
[5]AS = Argument Structure
[6]Please follow this link: https://clada-bg.eu/en/centers-and-services/language-technologies/services-and-tools.html and within it search for 'The integrated system for corpora and dictionaries'. The service is in Bulgarian.

FrameNet, was an advantage, since as mentioned above, we used the mappings between BTB-Wordnet and Princeton wordnet/English Open wordnet. The adaptation went into several directions such as: the number and names of the roles, the verb grammatical behavior, the treatment of metaphorical usages, etc. For the annotation task three annotators from Bulgarian Philology with good knowledge of English were selected as 2022 summer interns. Their task was to check the frames extracted from the treebank with respect to the verb meaning and available examples, as well as to edit, if necessary. Then, they had to assign the semantic roles to the syntactic arguments of each distinct verb having in mind the lexicographic class as a general pointer, and VerbNet with its verb frames for English. This semantic role annotation over the valency frames has had big influence on improving the coverage of BTB-Wordnet where the missing meanings and lemmas have been continuously added. Among the challenges during this process were these: the representation of the various types of multiword expressions (MWEs) and handling metaphorical usages.

3 Analyses

Let us first briefly introduce our notation of the valency frame. In Fig. 1 the valency frame of the verb 'celebrate' is given. We prefer to present this graphical view to the source XML for the sake of readability.

The frame is as follows:

```
1) Litse praznuvam sabitie
Person celebrate-1SG event
A person celebrates an event
```

Here the following information can be seen: the lexicographic wordnet class of *verb.social*, the LEMMA 'praznuvam' (celebrate-1P-SG-PRES), the definition (DEF) 'Have-I some holiday'.

The ones who celebrate, take the Agent role. The celebrated event is assigned the Theme role. The syntactic structure has been preserved. Here VPS means a phrase of type head-subject, and VPC means a phrase of type head-complement. On the graphics one cannot see the link to VerbNet[7] and the link to the English verb from the Open English Wordnet[8]. At the moment 28 semantic roles have been employed from VerbNet. We opted to have enough text examples and linguistic tests when assigning them.

Below come some examples with MWEs. The first one exemplifies an idiomatic expression while the second one – a light verb construction.

```
2) Igraya na kotka i mishka
Play-1SG on cat and mouse
I play cat and mouse
```

The semantic roles in example 2 are as follows: the player is an Agent (where the Co-Agent role is also presupposed) and then there are two levels of role assignment. On the first one, both pseudocomplements are a coordinated Theme to the predicate. On the second one, the predicate and its pseudocomplements are viewed as a synonym of the verb 'chase', thus only the Agent and Co-Agent roles remain active.

```
3) Podlagam nyakogo na stres
Make-1SG someone on stress
I am stressing someone out
```

The semantic roles in example 3 are as follows: The one who stresses somebody is a Stimulus, the stressed one is an Experiencer and the stress itself is a Theme. The lexicographic category of the construction is *verb.emotion*. Please note that this example allows for the compositional way of argument realization but this is not applicable elsewhere (e.g. take a shower).

In treating metaphors we follow the strategies in (Bonial et al., 2011 and Brown and Palmer, 2012). This means applying the set of roles for literal usages also to their metaphorical usages when possible.

[7] https://verbs.colorado.edu/verb-index/vn/judgment-33.php

[8] celebrate (wn 3; g 2)

For example, the Theme in VerbNet is set to 'a participant that is being literally or metaphorically located, positioned, or moved; this participant may be concrete or abstract.'

Table 1 presents the frequencies of the valency frames per lexicographic class.

Class	Freq	Class	Freq
verb.communication	736	verb.creation	188
verb.change	553	verb.competition	108
verb.cognition	500	verb.body	105
verb.motion	499	verb.consumption	77
verb.social	468	verb.weather	33
verb.stative	466	**Result**	**4736**
verb.contact	338		
verb.possession	258		
verb.perception	212		
verb.emotion	195		

Table 1: Frequencies of the valency frames per lexicographic class.

Since our valency frame assignment is corpus-based (mainly news media, partly literature and partly administrative texts), it can be seen that most frames go to *verb.communication* while the least ones – to *verb.weather*. Then come *verb.change* and em verb.cognition types – without big differences between them. Here we can classify also *verb.motion*. Interestingly, *verb.social* and *verb.stative* have nearly the same number of occurrences.

Our idea is to cross-check the verb semantic categories with the roles assigned within the distinct frames. With such cross-checks, validation can be made with respect to which verb groups get differing roles and what the reasons are behind the errors when using various resources like wordnets, VerbNet and argument linking theories; inconsistencies in VerbNet, wrong localisation to Bulgarian in knowledge transfer steps, specifics in the lexical semantics and context usages of the verbs etc.

Let us take as an example the *verb.perception* type. Here we should expect the prevalence of the default roles like Experiencer, Stimulus, Attribute. However, when a statistics was made over this type, the Theme (124 occurrences) turned out to be the most frequently used role, followed by Agent (72 occurrences) and immediately after it – by Experiencer (67 occurrences).Then the Stimulus comes with 50 occurrences as well as the Attribute with 27 occurrences. In the remaining list other diverse roles follow like Location, Destination, Co-Theme, Patient, Pivot, etc. When checking the examples, one can see that Agent comes in at least the following situations: a) the agent-like subject is an institution, not a person (for example, the government adopts some fiscal policy); b) the verb of perception includes control (watch, observe, look into vs. see or eavesdrop vs. hear) and c) verbs of producing sounds like roar, buzz.

4 Discussion and Conclusions

Since the verb semantic category is very important for the semantic roles assignment, we measured the number of the differing lexicographic classes between two annotators in their early stage of work. From 292 assignments 79 were different, which means more than one third. When we looked closer into the annotations, several issues appeared that were the source of this disagreement. The previous version of the valency dictionary did not have any semantic types added and the definitions adhered to the previous version of BTB-Wordnet. Also not so many examples were available to each lemma meaning. Thus, the annotators often had to match more general definitions to more fine-grained ones with not always having discriminatory contexts in examples or appropriate examples. In the differing cases they often opted for close to each other senses with differing types, or one of them selected a more specific one while the other - a more general one. Another issue was the handling of the multiword expressions. One selected the type of the non-MWE synonym while the other tried to get the type of the MWE predicate. Cross-resource and cross-language knowledge transfer is not a trivial task. In our future work we intend to look closer into these issues and make a strategy for handling them.

The main challenges in creating and then integrating various resources like valency dictionaries for Bulgarian are these: diversity in valencies within synsets (due to usage of a different preposition or no preposition); the idiosyncracy of MWEs in wordnets and valency dictionaries; the non-homogeneous behaviour of the reflexive verbs; the various aspectual nuances of verbs; the missing meaning and/or frame; the differing behaviour of the same verb in the two languages (where the same concept has been lexicalized); the asymmetries between the languages; the blurred boundaries among some semantic roles; the blurred boundaries among some senses; the handling of the optional arguments in the examples that have been attached to the respective senses, etc.

References

Birtić, M., Brač, I., & Runjaic, S. (2017). The Main Features of the e-Glava Online Valency Dictionary.

Bonial, C., Windisch Brown, S., Corvey, W., Palmer, M., Petukhova, V., & Bunt, H. (2011). An Exploratory Comparison of Thematic Roles in VerbNet and LIRICS [978-90-74029-35-3]. In H. Bunt (Ed.), *Proceedings of the 6th joint iso - acl sigsem workshop on interoperable semantic annotation* (pp. 39–43). University of Oxford.

Brown, S. W., & Palmer, M. (2012). Semantic annotation of metaphorical verbs: A Case Study of Climb and Poison. *Proceedings of ISA-8:Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*.

Di Fabio, A., Conia, S., & Navigli, R. (2019). VerbAtlas: A Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 627–637. https://doi.org/10.18653/v1/D19-1058

McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., & Fellbaum, C. (2019). English WordNet 2019 – An Open-Source WordNet for English. *Proceedings of the 10th Global Wordnet Conference*, 245–252. https://aclanthology.org/2019.gwc-1.31

Osenova, P. (2022). The Covid-19 Pandemic in the valency of Its Predicates: Observations on a Contemporary Corpus of Parliamentary Debates. *Bulgarian Language (Supplement)*, (69), 113–121.

Osenova, P., & Simov, K. (2018). The data-driven Bulgarian WordNet: BTBWN. *Cognitive Studies — Études cognitives*.

Osenova, P., Simov, K., Laskova, L., & Kancheva, S. (2012). A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. *International Conference on Language Resources and Evaluation*.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., & Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2785–2792. http://www.lrec-conf.org/proceedings/lrec2014/pdf/279_Paper.pdf

Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics*, *17*(4), 409–441. https://aclanthology.org/J91-4003

Pustejovsky, J., & Batiukova, O. (2019). *The Lexicon*. Cambridge University Press.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). *Syntactic Theory: A Formal Introduction*. Center for the Study of Language; Information.

Simov, K., Osenova, P., Simov, A., & Kouylekov, (2004). Design and Implementation of the Bulgarian HPSG-based Treebank. *Research on Language and Computation*, *2*, 495–522.

Straňáková-Lopatková, M., & Žabokrtský, Z. (2002). Valency dictionary of Czech Verbs: Complex Tectogrammatical Annotation. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. http://www.lrec-conf.org/proceedings/lrec2002/pdf/168.pdf

The ParlaMint Project: Ever-growing Family of Comparable and Interoperable Parliamentary Corpora

Maciej Ogrodniczuk
Institute of Computer Science PAS
Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Petya Osenova
Bulgarian Academy of Sciences
Sofia, Bulgaria
petya@bultreebank.org

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Darja Fišer
Institute of Contemporary History
Ljubljana, Slovenia
darja.fiser@inz.si

Nikola Ljubešić
Jožef Stefan Institute
Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Çagrı Çöltekin
Tuebingen University
Tuebingen, Germany
ccoltekin@sfs.uni-tuebingen.de

Matyáš Kopp
Charles University
Prague, Czech Republic
kopp@ufal.mff.cuni.cz

Katja Meden
Jožef Stefan Institute
Ljubljana, Slovenia
katja.meden@ijs.si

Taja Kuzman
Jožef Stefan Institute
Ljubljana, Slovenia
taja.kuzman@ijs.si

Abstract

We present an overview of the current status of the CLARIN flagship project ParlaMint (ParlaMint II), summarizing coverage of the parliament corpora, their size and time span. The organization of the workflow is outlined and newly added corpus enrichment procedures are described, such as machine translation of the corpora into English and showcasing of multimodal aspects. Last but not least, a number of ways for exploring the corpora are described.

1 Introduction

Debates in national parliaments are a verified communication channel between elected political representatives and society. They are also a reflection of the interests of the whole national community over time. This valuable data needs to be made accessible and comprehensible – especially in times of global crises. It should be possible to analyze such data synchronously and diachronously in a cross-lingual and cross-parliament context. However, parliamentary data per country/region exhibit various specific features. Each country/region has its specific type of parliamentary system; each parliamentary office publishes debates in its own way and the data comes in different formats, with different metadata and different structures. Thus, it is obvious that in their original form, these data cannot be compared or analyzed.

ParlaMint, a CLARIN ERIC-funded project,[1] manages to overcome the above mentioned problems with a harmonised representation format for parliamentary debates; compilation of a collection of parliamentary corpora aiming for 31 parliaments; their conversion to a common TEI format and their Universal Dependencies[2] oriented linguistic processing. Apart from making the data available for download and analysis, ParlaMint offers substantial help for adding new corpora to the dataset: encoding guidelines, validation procedures, documentation and samples.

ParlaMint is a solid data-intensive infrastructure, which can be used by various analytic tools to make parliamentary debates across Europe more transparent and comparable to researchers and citizens. We believe that it is just a starting point of a long-term impact action of bringing the accurate and trustworthy

[1] https://www.clarin.eu/parlamint
[2] https://universaldependencies.org/

information coming from our national parliaments to all parties interested in using it – and doing it in a cross-lingual perspective, which was never possible before.

The paper is structured as follows: in Section 2 the updated overview of the ParlaMint Corpora is presented. The focus in Section 3 goes beyond the creation of the corpora, thus introducing their machine translation into English and the multimodal perspectives of the data. Section 4 describes a plethora of ways to explore the resource. Section 5 concludes the paper.

2 ParlaMint corpora: an updated overview

ParlaMint was first funded by CLARIN ERIC during the COVID-19 pandemic. Since its first phase – ParlaMint I (July 2020 – May 2021) – was very successful with 17 parliamentary corpora compiled, processed and made available (Erjavec, Ogrodniczuk, et al., 2023),[3] the second phase was started – ParlaMint II (December 2021 – September 2023). In this second phase, 14 more parliaments are being added, including regional ones. We aim to cover the parliament sessions of the following parliaments: Austria, Basque Country, Bosnia and Herzegovina, Belgium, Bulgaria, Catalonia, Croatia, Czechia, Denmark, Estonia, Finland, France, Galicia, Greece, Hungary, Iceland, Italy, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovenia, Spain, Sweden, Turkey, UK, and Ukraine.

In Figure 1, an overview of ParlaMint corpora is given, distinguishing version 2.1 from version 3.0 (Erjavec, Kopp, et al., 2023), and the current status. It can be seen that they vary with respect to the time period and size of the data included. The corpora – with a few exceptions – have an overlapping period of transcripts, starting from 2015 and reaching to mid-2022. At the same time, some corpora provide data for longer pre-2015 periods dating back, such as Austria and Norway. The number of words does not depend only on the time span covered but also on the size of data provided and whether unicameral or bicameral proceedings are included.

Since one of the reasons for making the ParlaMint I corpora was the COVID-19 crisis, we marked the texts of the corpora as belonging to the 'covid' subcorpus or reference subcorpus, with the date marking the division between the two being the somewhat arbitrary 2019-11-01. In ParlaMint II, we changed this distinction to distinguish already three subcorpora: the 'war' subcorpus, starting at 2022-02-24, the date of Russia's full-scale invasion of Ukraine; the 'covid' subcorpus, starting at 2020-01-31, when WHO made the formal declaration of PHEIC (i.e., the Public Health Emergency of International Concern) for COVID-19; and the 'reference' corpus, until 2020-01-30. These distinctions are also indicated in Fig. 1.

The comparability among corpora reflects some other aspects as well: adherence to a specially designed common TEI encoding scheme and a validation procedure (discussed in the next section); to the availability of a same-language corpus (English) created from all the national corpora with machine translation techniques as well as of multimodal dimensions (outlined in subsection 3.1). The added value of the compiled corpora and accompanied procedures are presented in Section 4 through various ways of data usage and exploration.

3 Making ParlaMint corpora comparable and multimodal

One of the main goals of the ParlaMint project is to make the corpora of parliamentary debates comparable, i.e., compiling the corpora in several languages in a harmonised and uniform format. To achieve this aim, the ParlaMint I project's encoding process followed the TEI-based Parla-CLARIN recommendations (Erjavec & Pančur, 2021),[4] a flexible framework for encoding parliamentary corpora. As the project progressed, it became apparent that there was a need to further unify the corpus encoding and content, to make them more interoperable. To this end, a customised, more constrained RelaxNG scheme was developed and refined over the course of the ParlaMint I project, which allowed for more uniform encoding and supported corpus validation (Erjavec, Ogrodniczuk, et al., 2023; Ogrodniczuk et al., 2022).

[3] The release details of ParlaMint I corpora (i.e., version 2.1) can be found on ParlaMint webpage: https://www.clarin.eu/parlamint#links-to-current-parlamint-related-data-corpora-standards-servic.

[4] https://clarin-eric.github.io/parla-clarin/

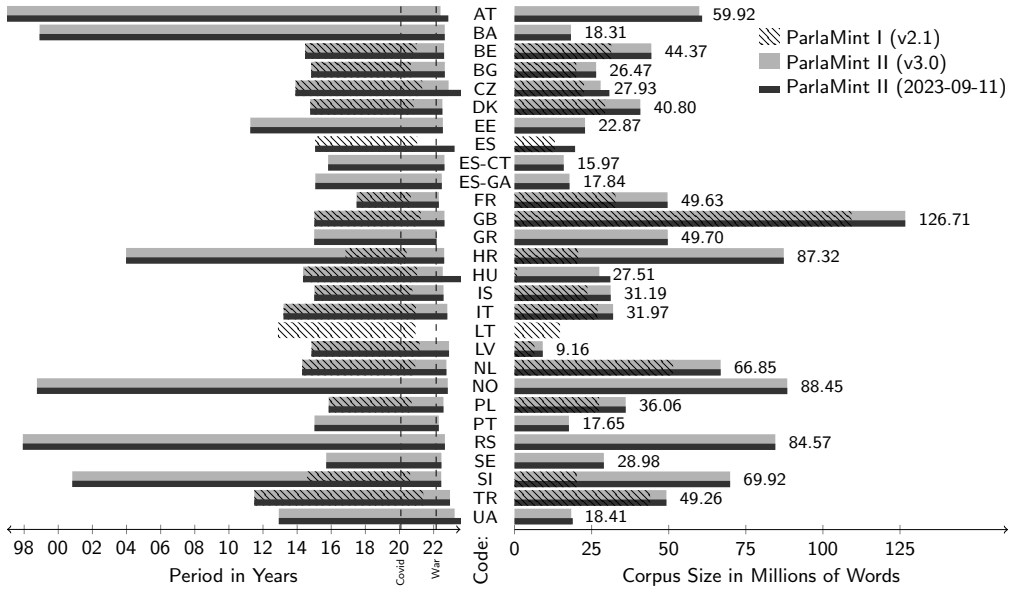


Figure 1: The time period and number of words of the ParlaMint corpora.

Following this line of reasoning, ParlaMint II first updated the original Parla-CLARIN guidelines to include solutions and encoding examples from the ParlaMint I corpora. Based on the updated guidelines, a set of new recommendations (XML TEI ODD document and prose guidelines[5]) was created to serve as a basis for encoding new corpora to be added or expanded in ParlaMint II. Similarly to the RelaxNG schemata in ParlaMint I, the encoding schema (and thus the recommendations) were further refined in accordance with input from project partners, documented in the project's GitHub Issues.

In addition to the comparability of the corpora at the encoding level, comparability was also facilitated at the metadata level. The original metadata included metadata on speakers and their political affiliation (name, gender, MP status, party affiliation, party coalition/opposition), while for the ParlaMint II we are adding two additional metadata labels: the political orientation of political parties (based on the Chapel Hill Expert Surveys dataset (CHES)[6] and Wikipedia, and metadata on which speakers are also ministers.

3.1 Machine Translation into English

To provide researchers with the opportunity to extend their research of parliamentary debates to all of the ParlaMint corpora, we translate them into English. Since the corpora are very large, consisting from 10 to 90 million words, human translation is impossible and we thus use machine translation. We use the pre-trained OPUS-MT models (Tiedemann & Thottingal, 2020), freely-available[7] Transformer-based models that are based on the MarianNMT neural machine translation toolbox (Junczys-Dowmunt et al., 2018) and trained on parallel corpora from the OPUS repository (Tiedemann, 2012). The great advantage of the OPUS-MT models is that they cover all of the ParlaMint languages. They are either specialized for a specific language, such as models for Polish, or for a language family, such as models for South Slavic languages, which we use where one-language models are not available. To choose the most appropriate model, a short manual evaluation of the translation output of the available models is performed for each language. The evaluation revealed that proper names are frequently incorrectly translated. Thus, in the corpora that use a Latin script and where the named-entity recognition works well, proper name translations are corrected by aligning words from the source sentence and translated sentence and substituting translated proper names with their lemmas from the source sentence. The machine translation consists of

[5]https://clarin-eric.github.io/ParlaMint/
[6]https://www.chesdata.eu/ches-europe
[7]https://github.com/Helsinki-NLP/Opus-MT

translating speeches from the CoNLL-U files and transcribers' notes from the TEI files. The translated corpora are linguistically processed with the Stanza pipeline (Qi et al., 2020) and made freely available as the ParlaMint-en.ana 3.0 corpora (Kuzman et al., 2023). They are connected with the original corpora and can be queried via concordancers[8] as a parallel corpus.

3.2 Multimodality Datasets

We are also preparing multimodal datasets, where speech recordings are aligned to the transcripts of parliamentary debates. Up to this point, a Croatian dataset of 1,816 hours of speech, named ParlaSpeech-HR, was published (Ljubešić, Koržinek, Rupnik, & Jazbec, 2022; Ljubešić, Koržinek, Rupnik, Jazbec, et al., 2022). This dataset is based on the previous iteration of the ParlaMint corpora, where Croatian was represented only through four years of parliamentary proceedings. Given that the current iteration of the ParlaMint corpus consists of 19 years of Croatian proceedings, our expectation is that the next iteration of the ParlaSpeech-HR dataset will be 4 times larger. We also plan to release similar datasets from the Polish, Czech, Bosnian and Serbian parliaments. Besides the data themselves, we plan to publish a robust pipeline for preparing comparable datasets in additional languages.

4 Exploring ParlaMint Corpora

In ParlaMint II, the engagement activities proliferated with the available 2.1 release and the upcoming new parliaments. ParlaMint data was used for the third time at DHH2023 hackathon, Helsinki (24 May–2 June 2023), with a topic on Political Polarization in the Parliament.[9] In 2021, the topic was related to parliamentary debates in covid times[10] and in 2022 the focus was on the parliamentary networks.[11]

There was a focused tutorial on how to use ParlaMint data in digital humanities at DH2023 Pre-conference, Graz (11 July 2023).[12] ParlaMint was also demonstrated at the CLARIN workshop, the CL2023 conference, Lancaster (2 July).[13] We also plan to run a shared task on multilingual ideology and 'language of power' detection on the ParlaMint corpora at CLEF.

ParlaMint data has inspired two impact stories so far. The first one is called "ParlaMint – A Resource for Democracy" (Dario Del Fante and Virginia Zorzi) and within it, the researchers explored how migration and migrants were referred to during the so-called migration crisis (2015/16) and the advent of COVID-19 (2020) in two countries – Italy and the UK. It was shown how this may impact public opinion on the topic.[14] The second one is called "Networks of Power – Gender Analysis in European Parliaments" (Jure Skubic, Alexandra Bruncrona, Jan Angermeier, Bojan Evkoski, Larissa Leiminger). The researchers worked with data from three parliaments – Slovenian, Spanish and British. The structural power within parliament was examined by applying speech selection to five topics: energy, finances, healthcare, education, and immigration. Here, the findings suggest that men hold more argumentative power than women in all studied parliaments, both in terms of Active Relevance and Passive Relevance.[15]

5 Conclusions and Future Work

ParlaMint has turned into a constant effort – emerged and ongoing in times of various health and social crises. With 31 parliaments planned, and a newly created subcorpus starting with the date of the Russian invasion on Ukraine, the project is offering a completely new pan-European parliamentary research landscape.

But several new research paths are still ahead. National datasets could also be compared with the data coming from the European Parliament and other regional parliaments. Parliamentary data is not just debates, but also voting results or the documents related to the law-making process. Integrating such data

[8] Available at: https://www.clarin.si/ske/, https://www.clarin.si/kontext/corpora/corplist and https://www.clarin.si/noske/.
[9] https://www.helsinki.fi/en/digital-humanities/dhh23-hackathon/dhh23-themes
[10] https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/
[11] https://www.helsinki.fi/en/digital-humanities/dh-h22-hackathon/dhh22-themes
[12] https://dh2023.adho.org/?page_id=616
[13] https://www.lancaster.ac.uk/cl2023/pre-conference-workshops/
[14] https://www.clarin.eu/impact-stories/parlamint-resource-democracy
[15] https://www.clarin.eu/impact-stories/networks-power-gender-analysis-european-parliaments

will give researchers and active citizens even more analytic possibilities. And there is a lot to be done with new and emerging technologies, focusing on processing multimodal data or producing live datasets, made available on the fly.

Acknowledgements

The work described in this paper was funded by CLARIN ERIC and by in-kind contributions of partners from ParlaMint I,[16] including POIR.04.02.00-00C002/19 and the Polish Ministry of Education and Science 2022/WK/09, programs Language resources and technologies for Slovene P6-0411 and Digital Humanities P6-0436, projects LiLaH N6-0099 and MEZZANINE J7-4642, all funded by the Slovenian Research Agency; CLaDA-BG DO01-301/17.12.21, funded by the Bulgarian Ministry of Education and Science; the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D., Pirker, H., Wissik, T., Schopper, D., Kirnbauer, M., Mochtak, M., Ljubešić, N., Rupnik, P., Pol, H. v. d., Depoorter, G., de Does, J., Simov, K., Grigorova, V., Grigorov, I., Jongejan, B., … Kryvenko, A. (2023). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1488

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T., & Pančur, A. (2021). The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.4133

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. *Proceedings of ACL 2018, System Demonstrations*, 116–121. http://www.aclweb.org/anthology/P18-4020

Kuzman, T., Ljubešić, N., Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D., Pirker, H., Wissik, T., Schopper, D., Kirnbauer, M., Mochtak, M., Rupnik, P., Pol, H. v. d., Depoorter, G., de Does, J., Simov, K., Grigorova, V., Grigorov, I., … Kryvenko, A. (2023). Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 3.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1810

Ljubešić, N., Koržinek, D., Rupnik, P., & Jazbec, I.-P. (2022). ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 111–116.

Ljubešić, N., Koržinek, D., Rupnik, P., Jazbec, I.-P., Batanović, V., Bajčetić, L., & Evkoski, B. (2022). ASR training dataset for Croatian ParlaSpeech-HR v1. 0 [http://hdl.handle.net/11356/1494].

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M., & Katja, M. (2022). ParlaMint II: The Show Must Go On. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 1–6.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *LREC*, *2012*, 2214–2218.

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT–Building open translation services for the World. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

[16] https://www.clarin.eu/parlamint#Partners

Workflow and Metadata Challenges in the ParlaMint Project: Insights from Building the ParlaMint-UA Corpus

Anna Kryvenko
Institute of Contemporary History, Slovenia;
NISS, Ukraine
Ganna.Kryvenkol@inz.si

Matyáš Kopp
Charles University
Prague, Czech Republic
kopp@ufal.mff.cuni.cz

Abstract

This paper focuses on the challenges of refining the workflow for collecting and adding metadata to the ParlaMint corpora designed for research in the social sciences and humanities. The ParlaMint project aims to create a multilingual family of comparable and uniformly annotated corpora containing parliamentary proceedings from European national and regional parliaments. In particular, we report on a workflow for the automated and manual processes of collecting and adding metadata to the newly minted ParlaMint-UA corpus based on the ParlaMint TEI schema for corpora of parliamentary proceedings. We also specify some categories related to legislative periods, speaker roles, party affiliations and political relations in the corpus data. We argue that these findings may contribute to best practices for the construction of politically marked corpora.

1 Introduction

The ParlaMint project[1] (Erjavec, Ogrodniczuk, et al., 2023), now a multilingual family of 26 individual corpora ParlaMint 3.0 (Erjavec, Kopp, et al., 2023) containing parliamentary proceedings from European national and regional parliaments developed under the auspices of CLARIN-ERIC, has always aimed at building comparable and uniformly annotated corpora, which would cover the mandatory period of plenary meetings between 2015 and 2022 and be enriched with speaker and party metadata (Ogrodniczuk et al., 2023). However, the compilation of individual corpora for the ParlaMint project posed specific challenges primarily depending on whether the data was taken from existing corpora or was prepared from scratch, whether the metadata was available for automatic retrieval as well as whether the new corpora contained new phenomena to be encoded. At times, the latter led to "revisions of already accepted encoding practices, and hence to revisions of previously completed corpora" (ibid.). In this paper, we report on workflow adjustments for building the Ukrainian corpus for the project. We also discuss the importance of refining definitions for some metadata categories by providing more specific information about the types of referents that fall into these categories. We believe that our findings may be applicable to other ParlaMint corpora, especially those, where the speaker–speech affiliation is not encoded in the source, and more broadly contribute to best practices for the development of parliamentary corpora.

The ParlaMint-UA corpus is the first Ukrainian parliamentary corpus (Kryvenko, 2018), which was compiled and made comparable with the other national and regional European corpora within the ParlaMint project, although the original plans did not include the Ukrainian data (Ogrodniczuk et al., 2022). It contains records of plenary proceedings from the Verkhovna Rada[2] – the unicameral parliament of Ukraine – for Terms 7, 8 and 9 between 12 December 2012 and 24 February 2023. Archived transcripts of all plenary sittings were acquired through the open data portal[3] at the Rada site in HTML format under the CC BY 4.0 license. The total numbers found in the ParlaMint-UA 3.0 corpus include over 22.5M tokens, 18M words, 1.5M sentences and 195k speeches. Language identification is done at the paragraph[4]

[1] https://www.clarin.eu/parlamint#parlamint-ii-work-plan

[2] https://www.rada.gov.ua/meeting/stenogr

[3] https://data.rada.gov.ua/open

[4] The corpus preserves paragraph segmentation produced by stenographers in the Rada.

level, with 99 % of utterances in Ukrainian and 1 % in Russian. The corpus embraces 783 sitting dates, 1,475 speakers and 52 parliamentary parties, factions, and groups.

Contrary to other national parliaments, such as Czech with the ParCzech project (Kopp et al., 2021), Ukrainian official sources do not preserve historical metadata very well, so various sources had to be used to get required personal and organizational metadata. The metadata related to MPs were in part retrieved from the Rada website and in part gathered manually from official sources, including the Central Election Commission of Ukraine, the official periodical of the Rada, Holos Ukrainy, the Rules of Order of the Rada and other open data sources. The metadata related to Cabinet members and guest speakers were gathered manually from the current sites of the Cabinet of Ministers of Ukraine and the Rada, archived copies of webpages from the sites of the Rada, the Cabinet, the President of Ukraine, and various open data sources including NGOs' websites, mass and social media as well as Wikipedia.

In what follows, we describe the workflow we developed to integrate automatic and manual collection and the adding of the metadata on the speakers, organizations and events involved, since there was no single source containing all the information required for a uniform annotation of the corpus under the ParlaMint TEI schema. Then, we focus on some categories of speaker and party metadata, which required a more refined definition of the scope of their notion and the circle of their referents to be applied consistently across the corpus. Finally, we draw conclusions and suggest applications of our findings.

2 Workflow

The workflow[5] for collecting and adding the metadata on speakers, organizations and events for Terms 7–9 included 11 main steps, both single and repetitive, as shown in Figure 1. During the initial stage, plenary transcripts, lists of parliamentary speeches containing timestamps, and personal metadata on MPs, including their full names, dates of birth, gender, and affiliations within the Rada, were automatically downloaded in the HTML, XML and CSV formats from the Rada open data portal. The metadata on government members, guest speakers, organizations and events like the periods of governments in office, as well as additional metadata on MPs like person renaming, were collected manually from open sources and organized as a spreadsheet with the following: 1) conditional formatting for more accessible validation of the filled data; 2) automatic generation of IDs; 3) publishing the spreadsheet with the "publish to the web" Google Sheet feature, so the tables could be downloaded in the tsv file format with the script.

In the meantime, the textual part of the TEI format was produced. The HTML proceedings were converted to the ParlaMint TEI format by segmenting input into utterances and paragraphs, categorizing paragraphs by language, and annotating nonspeech content such as interruptions and notes. In the next step, the files were morphologically and syntactically annotated according to the Universal Dependencies formalism, then named entities were recognized and the results were stored under the TEI.ana label.

The TEI.ana files, along with the lists of speeches and the speaker and party metadata, were utilized for the automatic linking between speakers and persons. When this linking failed, the TEI.ana files were used to automatically detect mentions of the mismatching or multimatching speakers in the preceding utterances, commonly made by chairpersons. Detecting mentions was an important step toward improving the speakers-persons linking, because it helped to: 1) get full names of speakers and therefore reduce the amount of manual work; 2) disambiguate between the speakers with homonymous surnames and abbreviated forenames and patronymics in the transcripts; 3) handle typos in the speakers' names, for which we used the Levenshtein distance of 3 edits at maximum.

The retrieved lists of mis(multi)matching speakers were aligned with the manually collected metadata in the spreadsheet. The remaining mis(multi)matching speakers were linked manually before the ParlaMint-UA corpus was finalized. Keeping the speaker–speech links apart and merging them at the end of the pipeline made it possible to fix any bugs related to those links without reprocessing part of the corpus. This practice is applicable when the speaker–speech affiliation is not encoded in the source, i.e., there are no unique speaker IDs in the source and the linking is done based on the textual form of the name. The corpus finalization wound up all the other processes like setting header information in every TEI file, affiliating persons' IDs to utterances, and calculating volumes of elements, words and speeches.

[5] All the scripts used in the project are hosted on GitHub: https://github.com/ufal/ParlaMint-UA

It is also important to emphasise that the Ukrainian team consisted of only two members: the programmer who was in charge of all automated tasks required for compiling and submitting the Ukrainian corpus to the ParlaMint repository, and a digital humanities specialist without knowledge of programming or the XML file format who took care of all the manual tasks including editing the dedicated Google Sheet that reported bugs with conditional formatting. This practice can be suitable for new ParlaMint contributors in terms of optimizing labour efficiency and reducing errors in the corpus metadata.

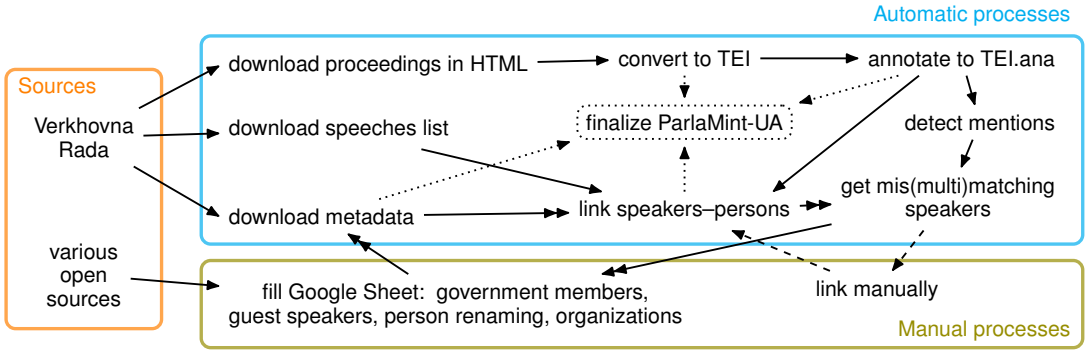


Figure 1: Workflow for collecting and adding metadata to the ParlaMint-UA corpus

3 Definitions of metadata categories

A TEI Schema for Corpora of Parliamentary Proceedings [6] provides explanations of metadata categories in the example documentation. However, while building the ParlaMint-UA corpus, we came to the realization that the categories including but potentially not limited to the parliamentary term (typically given as `<event>`), chairperson, regular speaker, guest speaker (all three are values of speaker role), acting minister (a new attribute of `<affiliation>`), and opposition and coalition (stored in `<relation>` elements) need to be further defined with respect to variation of the underlying phenomena in the parliaments. In what follows, we offer our definitions of the terms above for the purposes of the ParlaMint-UA corpus.

A parliamentary term, or legislative period, or convocation, is the period between the day of the first sitting of the current term and the day of the first sitting of the next term. MPs' mandates are in power until the opening of the first sitting of the next parliament according to the Constitution of Ukraine, although early terminations are possible.

A chairperson is an MP presiding in a sitting, who the Rada elected as the Chairperson, the First Deputy Chairperson, Deputy Chairperson, or the Chairperson of the Preparatory Parliamentary Group, who is elected as a presiding officer prior to the election of the Chairperson of the Rada at the beginning of the new legislative period.

Regular speakers include: 1) MPs who do not preside at plenary sittings; and 2) Members of the Cabinet of Ministers as well as deputy ministers, neither of whom are MPs under Ukrainian laws but who regularly participate in the sittings during the "hour of questions to the Government" according to the Rules of Order of the Rada. The practice of including non-presiding MPs and ministers into the same category stems from those parliaments in the ParlaMint project where members of the government are allowed to simultaneously hold their MP mandate (e.g., the Czech parliament). However, unlike in the Czech parliament, it is common that Ukrainian deputy ministers can represent their ministries in the Rada, although they are not Members of the Cabinet of Ministers of Ukraine.

Guest speakers are a diverse category of plenary sitting participants in the Rada including the current and the former presidents of Ukraine, former MPs, representatives of central authorities other than the Cabinet of Ministers as well as local governments, religious leaders, civil activists, and foreign politicians. They can be invited to either ceremonial or regular plenary sittings. Furthermore, the President of Ukraine is bound to deliver annual statements to the Rada.

An acting minister is a newly added label for the affiliation role (along with ministers and deputy ministers). It is used in the ParlaMint-UA corpus for government officials who were appointed to serve

[6] https://clarin-eric.github.io/parla-clarin/

in the role of a minister or a deputy minister on an interim basis but not to hold a respective office. We are not aware of the usage of this extended affiliation role by any other corpora in the ParlaMint project.

To distinguish between the coalition and the opposition in the Rada, we used the following indicators: participation of the parliamentary groups in the formation of the Cabinet of Ministers of Ukraine (a total of six governments were installed in the course of the last three parliamentary terms); announcements by the leaders of parliamentary groups about switching to the opposition at the plenary sittings of the Rada; voting for key laws like the Annual Budget or key appointments like the Head of the Security Service of Ukraine after the large-scale aggression of Russia against Ukraine in February 2022.

4 Conclusion

In this contribution, we presented the workflow we had developed to integrate automatic and manual metadata collection and inclusion to the ParlaMint-UA corpus. This workflow enhanced the process of linking speakers and persons as well as the finalization of the corpus.

We also reported on refining the definitions of names for some key metadata categories, which we elaborated based on the Ukrainian legislation and practices in the Verkhovna Rada of Ukraine. We encourage other parliamentary corpus teams to include in their corpus documentation statements about the scope of notions and the circle of referents applicable to the metadata categories of speakers and their organizations, which are specific to their parliaments. Addressing these specificities is a step toward even greater comparability of the ParlaMint corpora, as it will allow researchers from different fields to better differentiate between variation in the transcripts and variation in the parliamentary systems involved.

Acknowledgments

The ParlaMint-UA corpus was developed with support from the program Digital Humanities P6-0436 by the Slovenian Research Agency and project N6-0288 by the Slovenian Research Agency, and CLARIN ERIC project 'ParlaMint: Towards Comparable Parliamentary Corpora'. Also, it used tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Fišer, D., Pirker, H., Wissik, T., Schopper, D., Kirnbauer, M., Mochtak, M., Ljubešić, N., Rupnik, P., Pol, H. v. d., Depoorter, G., de Does, J., Simov, K., Grigorova, V., Grigorov, I., Jongejan, B., … Kryvenko, A. (2023). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 3.0. http://hdl.handle.net/11356/1488

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Kopp, M., Stankov, V., Krůza, J. O., Straňák, P., & Bojar, O. (2021). ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata. In K. Ekštein, F. Pártl, & M. Konopík (Eds.), *TSD* (pp. 293–304). Springer International Publishing. https://doi.org/10.1007/978-3-030-83527-9_25

Kryvenko, A. (2018). Constructing a Narrative of European Integration in the Verkhovna Rada of Ukraine: A Corpus-Based Discourse Analysis. *Cognition, communication, discourse*, *17*, 56–74. https://doi.org/10.26565/2218-2926-2018-17-04

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M., & Katja, M. (2022). ParlaMint II: The Show Must Go On. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 1–6.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M., Meden, K., & Kuzman, T. (2023). The ParlaMint Project: Ever-growing Family of Comparable and Interoperable Parliamentary Corpora. *Proceedings of the CLARIN 2023 annual conference*, In print.

Adding political orientation metadata to ParlaMint corpora

Tomaž Erjavec
Department of Knowledge Technologies,
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Katja Meden
Dept. of Knowledge Technologies,
Jožef Stefan International Postgraduate School,
Jožef Stefan Institute, Slovenia
katja.meden@ijs.si

Jure Skubic
Institute of Contemporary History,
Ljubljana, Slovenia
jure.skubic@inz.si

Abstract

The speeches in ParlaMint corpora of parliamentary proceedings are marked by their speaker, and the speakers are then paired with various metadata, also with their time-delimited affiliations with political parties or parliamentary groups. These are stored separately, and are also associated with further information. This paper discusses the addition of metadata on political parties and parliamentary groups, encoding their political position on various issues, in particular their categorisation on the left-to-right political spectrum. The paper explains our sources for this information, the process of data collection, and its encoding in the corpora. This additional metadata should be of interest to parliamentary data research, while the methodology developed could be used to add further metadata to the ParlaMint corpora.

1 Introduction

The ParlaMint[1] project, funded by CLARIN, aims to create comparable and uniformly encoded corpora of speeches in European parliaments and make them openly accessible. Under ParlaMint I (2020-2021), corpora for 17 European parliaments were created, made available, and used in research and educattion (Erjavec et al., 2022). The project continued as ParlaMint II (2022-2023), providing 9 new corpora, adding newer transcripts, improving the annotation schema and validation, machine-translating the corpora into English, and expanding the corpus metadata.

Additional metadata added to the corpora consists of whether and when a speaker is or was a minister, and the political orientation of the parliamentary group or political party to which the speaker belongs. Both of these additions have been suggested by researchers (cf. Fišer and Pahor De Maiti, 2021) who had experience in using ParlaMint corpora, so that analyses could take these further variables into account. But while the information on who is a minister is an objective and verifiable fact that can be easily found, political orientation is a much more controversial piece of information.

2 Political orientation

Political orientations (or political positions) are an interesting research concept in the social sciences, understood as a set of ethical ideals, principles, or doctrines of a social movement, institution, class, or large group that explain how society should function and provide a political and cultural blueprint for a particular social order (Blattberg, 2001). They are concerned with the allocation and use of (political) power and are usually pursued by political parties.

Political orientation can refer to any number of dimensions but is most often characterized and classified on a political left-to-right spectrum, usually represented with geometric axes corresponding to independent political dimensions (Heywood, 2021). The left-to-right (LR) dimension is one of the most

[1] https://www.clarin.eu/parlamint

common dimensions used as a measure of social, political, and economic stance. Originally, the terms ”left” and ”right” were used to describe the nature and ideological beliefs of political parties: ”left” as the ”parties of movement,” which are radical, progressive, and liberal, and ”right” as the ”parties of order,” which are conservative, traditional, and authoritarian (Knapp & Wright, 2006), and such classification has, although in various forms, been retained until today. Left-right conceptualization is often considered controversial not only in terms of being defined as too simplistic and insufficiently representative to describe variations in political beliefs but also in terms of dimensionality. Most commonly LR divide is understood as unidimensional (structured by socio-economic issues) whereas some authors opt for multidimensionality where despite the importance of the socio-economic content, the left-right divide also correlates with other, non-economic issues (such as religious or ”new politics” issues) (Freire, 2015). Despite said controversies, left-right conceptualization is still the most common way to describe the ideological position of political parties and their members. The division into ”left” and ”right” has formed a categorization of ideologies, a tool for classifying political orientation, a communication code, and an instrument for guiding voters in interpreting decisions and political phenomena (Freire, 2015).

The left-right characterization of political parties plays a crucial role in theorizing about many different aspects of democratic processes (Gabel & Huber, 2000), and sociology and political science have adopted and used it despite various scholarly reservations. Some disciplines, such as history, however, often refrain from using the left-right political spectrum to characterize the ideological beliefs of political parties.

Most work in NLP attempts to determine political orientation directly from texts (whether from political tweets (Cohen & Ruths, 2021) or parliamentary debates (Yan et al., 2017)) and thus focuses on individual speeches. Unlike related work, we have instead focused on providing information about the political orientation of a political party rather than speeches, and thus take the political orientation of a speech to follow from membership in a particular party to which the speaker belongs at the time of his speech.

3 Data sources

The information on the political orientation of political parties contained in the ParlaMint corpora was gathered from three sources: (1) the Chapel Hill Expert Survey Europe (CHES Europe) (Jolly et al., 2022)[2]; (2) Wikipedia entries on political parties; and (3) the corpus compilers' knowledge of political parties and their orientations. We discuss each one in turn.

Chapel Hill survey: The CHES datasets contain expert data with built-in contextual and domain knowledge. They contain data on parliamentary political parties from countries, primarily from the EU, their attitudes toward European integration and specific EU policies, and on more specific topics such as corruption and anti-Islam rhetoric. We used two CSV files provided by CHES, namely the 1999-2019 trend file[3], which gives the values of the variables according to the covered years, and CHES 2019[4], which adds data for Norway, Iceland, and Turkey, as these were not covered in the CHES 1999-2019 trend file. This also means that these three corpora do not contain diachronic data.

The union of both CHES files provides 85 distinct variables on a given (political) position for each party and year covered, with most having a real value on the scale from 0 to 10, e.g. the variable LRGEN measures the party's position in relation to its overall ideological stance on a scale from 0 (extreme left) to 10 (extreme right), with 5 representing the centre position. This wealth of data could be of great value to political scientists basing their research on the ParlaMint corpora. However, the CHES information also has drawbacks which can be seen especially in its coverage:

- CHES does not cover all ParlaMint corpora, in particular Bosnia, Serbia and Ukraine, as they are not part of the EU (candidate countries), nor Catalonia and Galicia, as they are not countries but autonomous regions;

[2]https://www.chesdata.eu/ches-europe
[3]https://www.chesdata.eu/s/1999-2019_CHES_dataset_meansv3.csv
[4]https://www.chesdata.eu/s/CHES2019V3.csv

- Many political parties included in ParlaMint could not be identified in the CHES dataset: of the 576 political parties belonging to the countries covered by CHES and that are included in ParlaMint, only 237 (41%) could be matched with a CHES party identifier;
- Even for the parties that are identified, CHES only covers the period to 2019, while ParlaMint extends to 2022; furthermore, not all variables are covered for all years, nor do the two input files share all the variables.

Wikipedia: The second source and type of data included is Wikipedia, in particular the data on the left-right spectrum of political orientation. This data was gathered by manually searching for the political parties' Wikipedia pages, which typically list their political orientation in the infobox of the Wikipedia article, although for some, a more detailed examination of the Wikipedia article was required. We based our research on the English versions of the Wikipedia pages. When we could not find relevant information on the English page, we searched and translated the Wikipedia pages in the native language of the party's country.

Wikipedia uses values ranging from far-left to far-right, as well as 5 additional values which refer to specific political orientations outside the left-right scope. In total, we identified 13 different values which are shown in Table 1.

Abbreviation	Value
FL	Far-left
LLF	Left to far-left
L	**Left**
CLL	Centre-left to left
CL	Centre-left
CCL	Centre to centre-left
C	**Centre**
CCR	Centre to centre-right
CR	Centre-right
CRR	Centre-right to right
R	**Right**
RRF	Right to far-right
FR	Far-right
BT	**Big tent**[5]
PP	**Pirate Party**[6]
SY	**Syncretic politics**[7]
SI	**Single-issue politics**[8]
NP	**Nonpartisanism**[9]

Table 1: Political orientation values, identified in the Wikipedia data.

The information from Wikipedia covers the ParlaMint political parties and parliamentary groups quite well: out of 932 such entities currently defined in ParlaMint, only 20 (2.2%) could not be assigned a left-right orientation.

[5] A big tent party, or catch-all party, is a term used in reference to a political party's policy of permitting or encouraging a broad spectrum of views among its members. https://en.wikipedia.org/wiki/Big_tent.

[6] Pirate Party refers to political parties that support civil rights, direct democracy, encourage innovation and creativity, free sharing of knowledge, information privacy, free speech, anti-corruption, net neutrality and oppose mass surveillance, censorship and Big Tech. https://en.wikipedia.org/wiki/Pirate_Party.

[7] Syncretic politics refers to politics that combine elements from across the conventional left–right political spectrum. https://en.wikipedia.org/wiki/Syncretic_politics.

[8] Single-issue politics refers to a political stance that is based on one essential policy area or idea.https://en.wikipedia.org/wiki/Single-issue_politics.

[9] Nonpartisanism refers to a political stance that does not agree with the current political party system. https://en.wikipedia.org/wiki/Nonpartisanism.

Encoder classification: The third source of data was the encoders (i.e. compilers of the corpus), who, if they so decided, entered their classification on the left-right orientation, which was mainly to be able to mark the political parties that were not covered by Wikipedia. Currently only three of the partners made use of this option.

4 Data encoding

ParlaMint corpora are encoded in XML following the Text Encoding Initiative (TEI) Guidelines. The structures encoding the added metadata can be rather complex, so, in order to simplify the process of adding metadata and make it less error-prone, we did not require the orientation data to be entered directly into XML, but rather prepared tabular TSV files for each country that were pre-populated with the abbreviations of all the political parties. The Wikipedia URLs and orientation data, as well as the encoder orientation data were then added in Excel, possibly with comments, and the files saved as TSV[10]. An XSLT script then takes the TSV files and the XML corpus file with organisational data and inserts the new data into the XML. A similar procedure was applied to the CHES data: here the CHES CSV files were also converted to TSV, the party abbreviations from CHES semi-automatically mapped to the ParlaMint party identifiers in Excel, the results saved as TSV, and, again, inserted into the XML files.

```
<org role="parliamentaryGroup" xml:id="MR">
  <orgName full="abb">MR</orgName>
  <orgName full="yes">Mouvement Réformateur</orgName>
  <idno type="URI"
  subtype="wikimedia">https://en.wikipedia.org/wiki/Reformist_Movement</idno>
  <state type="politicalOrientation">
    <state type="encoder" source="#GrietDepoorter" ana="#orientation.CRR">
       <note xml:lang="en">Orientation determined by encoder, using own
       knowledge of the parliamentary group.</note>
    </state>
    <state type="Wikipedia"
     source="https://en.wikipedia.org/wiki/Reformist_Movement"  ana="#orientation.CR">
      <note xml:lang="en">From 1992 the Reformist Movement (MR) consisted of: FDF, MCC,
      PRL and PFF. In September 2001, FDF decides to leave the alliance and chooses a
      new name, becoming DeFI.</note>
    </state>
  </state>
  <state type="CHES" source="https://www.chesdata.eu/s/1999-2019_CHES_dataset_meansv3.csv">
    <label>
       <orgName full="abb" xml:lang="en" from="2002" to="2018">MR</orgName>
    </label>
    <state type="variable" ana="#ches.lrgen">
      <state type="value" from="2002" to="2005" n="6.35"/>
      <state type="value" from="2006" to="2009" n="6.67"/>
      <state type="value" from="2010" to="2013" n="7.0"/>
      <state type="value" from="2014" to="2018" n="7.0"/>
    </state>
    ...
    <state type="variable" ana="#ches.vote">
      <state type="value" from="2002" to="2005" n="10.1"/>
      <state type="value" from="2006" to="2009" n="11.4"/>
      <state type="value" from="2010" to="2013" n="9.28"/>
      <state type="value" from="2014" to="2018" n="9.6"/>
    </state>
  </state>
</org>
```

Figure 1: Encoding of political orientation in ParlaMint.

Figure 1 gives an example of the political orientation encoding. It should be noted that the CHES variables as well as the Wikipedia and encoder left-right orientations are pointers to taxonomy categories,

[10]The TSV files are available on the ParlaMint GitHub page, at https://github.com/clarin-eric/ParlaMint/tree/main/Corpora/Orientations

which give the name and explanation of the reference, e.g. similarly to the categories and explanations presented in Table 1.

5 Conclusions

We presented ongoing work to add political orientation metadata to the ParlaMint II parliamentary corpora. We have captured the political orientation of more than 350 European political parties by relying on two highly informative data sources, the Chapel Hill Expert Survey dataset and the Wikipedia pages of the respective parties.

We are aware that the political orientation of parties does not necessarily coincide with the personal orientation of the speaker belonging to the respective party and also recognize that people's ideological beliefs, as well as what they say, are often fluid and therefore difficult to capture. Nevertheless, the method that we have employed does give each speech its implied political orientation, and it should be interesting to attempt analyses based on this data. In the future, we would also like to use the metadata to gain additional insight into the data by analysing individual corpora using the CHES variables. Specifically for political orientation, we would also like to enable comparison of the speeches of left/right or centre-leaning speakers (or political parties) with each other to see if they speak in accordance with their political orientation. This type of comparison could then also be done for specific topics (e.g., attitudes toward European integration) that are included in the CHES metadata. In addition, the metadata will be used as part of the shared task on ideology and power identification in parliamentary debates[11], which will be part of the Touché lab[12] at the CLEF 2024[13] conference[14].

References

Blattberg, C. (2001). Political philosophies and political ideologies. *Public Affairs Quarterly*, *15*(3), 193–217.

Cohen, R., & Ruths, D. (2021). Classifying Political Orientation on Twitter: It's Not Easy! *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 91–99. https://doi.org/10.1609/icwsm.v7i1.14434

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings [https://doi.org/10.1007/s10579-021-09574-0]. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0

Fišer, D., & Pahor De Maiti, K. (2021). "First, I'm a Female Politician, Not a Male One, and Second …": A Corpus Approach to Parliamentary Discourse Research. *Contributions of contemporary history*, *61*(1), 144–179. https://doi.org/10.51663/pnz.61.1.07

Freire, A. (2015). Left–right ideology as a dimension of identification and of competition. *Journal of Political Ideologies*, *20*(1), 43–68.

Gabel, M. J., & Huber, J. D. (2000). Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science*, 94–103.

Heywood, A. (2021). *Political ideologies: An introduction*. Bloomsbury Publishing.

Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, *75*, 102420. https://doi.org/https://doi.org/10.1016/j.electstud.2021.102420

Knapp, A., & Wright, V. (2006). *The government and politics of france*. Routledge.

Yan, H., Lavoie, A., & Das, S. (2017). The perils of classifying political orientation from text. *Linked Democracy: Artificial Intelligence for Democratic Innovation*, *858*, 8.

[11] https://touche.webis.de/clef24/touche24-web/ideology-and-power-identification-in-parliamentary-debates.html

[12] https://touche.webis.de/clef24/touche24-web/index.html

[13] http://clef2024.clef-initiative.eu/

[14] For simplicity, only the left-to-right labels will be used, flattening the fine-grained annotations but still making use of the metadata.

MATEO: Machine Translation Evaluation for Users and Developers

Bram Vanroy
Ghent University; KU Leuven (current)
bram.vanroy@kuleuven.be

Abstract

This paper introduces MAchine Translation Evaluation Online (MATEO), a CLARIN *Bridging Gaps* project that simplifies the process of evaluating machine translation (MT) by providing an intuitive and user-friendly web interface that caters to both novice and expert users. It is equipped with a comprehensive suite of automatic metrics that can assess the quality of given machine translations by means of reference translations. The MATEO project serves a wide range of users including individuals with varying levels of experience in MT, such as system builders, educators, students, and researchers in (machine) translation as well as the social sciences and humanities. The interface is integrated in the CLARIN infrastructure and instructions are available for users to run the tool on their own device or in the cloud for free. It is open-source and GPLv3 licensed.

1 Introduction

The rapid development of machine translation (MT) in recent years has gone hand in hand with the creation of appropriate machine translation evaluation methods. As MT has gotten closer to human translation in terms of accuracy and fluency, so have quality evaluation metrics achieved higher correlation with human quality judgements. These evaluation metrics range from lexical or character-based matching statistics between a machine translation and a reference translation, to measures making use of recent large language models. While the performance of these metrics is indeed increasing, as is evident from the yearly shared task on machine translation metrics (WMT Metrics Shared Task; Freitag et al., 2022),[1] some of them are scattered across different GitHub repositories – if they are even publicly available – and they are not standardised in terms of implementation and usage. For technical users such as MT builders that is inconvenient: to evaluate their MT system with a number of evaluation metrics, they need to collect, install and run the different programs to calculate evaluation measures. This can be tedious and time-consuming but, what is worse, small differences in usage between metrics can lead to involuntary misuse and unexpected results. This project aims to address this problem by aggregating common and state-of-the-art evaluation metrics into a single Python-based web interface that is centred around the user and the developer. It focuses on the user by making it straightforward to calculate multiple metrics at the same time for a given dataset, and it emphasises the developer by specifically providing an open-source, extensible code base where new metrics can be added with ease.

Machine translation evaluation is not a trivial task, as it involves choosing appropriate metrics, obtaining reference translations, and analysing the results. While some of these steps are part of the workflow of the technical users addressed above, they are daunting for another group of stakeholders. Educators, students, and researchers from social sciences and humanities (SSH) and beyond, including translation scholars, and researchers in the domains of digital humanities and (computational) social sciences and political studies, can benefit from using machine translation evaluation for their purposes, such as comparing different translation systems and their impact on meaning, style and cultural context, or to evaluate in-domain machine translations. Technical requirements such as installing software and working on the

[1] https://wmt-metrics-task.github.io/

command line can be intimidating however. Therefore, this project delivers a web interface that facilitates the evaluation process. Such an interface has a number of benefits to involve these users in machine translation evaluation. It saves time and resources by avoiding the need to install and run complex software or scripts for MT evaluation and it provides a standardised and transparent way of conducting MT evaluation, which can enhance the reliability and reproducibility of the results for research and education. Importantly, a web interface also creates room to provide informative sections on machine translation evaluation, improving the digital literacy of this second group of users.

The MATEO project has been designed to address the needs of these two main groups: expert users of machine translation evaluation who have prior technical experience with MT evaluation, and a general audience consisting of, among others, students and educators, and researchers from SSH domains. MATEO is open-source and GPLv3 licensed.[2] The web interface is incorporated in the CLARIN B Center of the Dutch Language Institute (Instituut voor de Nederlandse Taal) to ensure the broad availability of machine translation evaluation[3], and, for the sake of longevity and openness, instructions are also provided to run it on users' own devices by means of Python or Docker. For less technical users, a guide is provided on how to host a private instance in the cloud on the Hugging Face Spaces infrastructure.

2 Related work

MATEO brings together different MT evaluation metrics in one interface that can be used by expert and non-experts, and it is built with Python so that developers can make preferential adjustments or add new metrics. Similar developments exist but have different points of focus or do not contain the functionality or metrics that MATEO wants to provide. Here I will discuss two types of related tools: programming frameworks that are intended to be used by expert users on the command-line or in Python, and web interfaces that are more generally usable.

2.1 Python frameworks

Most notably in the field of machine translation evaluation, `sacrebleu` (Post, 2018) offers users the ability to easily calculate a few common surface metrics BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and ChrF (Popović, 2017). As such, state-of-the-art metrics are not included but MATEO makes uses of `sacrebleu` behind the scenes to calculate these three metrics. Both a Python and command-line interface are provided and multiple systems can be statistically compared at the same time. It was an important development in MT evaluation to improve the standardization of evaluation metrics.

TorchMetrics by LightningAI is a general-purpose framework that is built with *training* deep learning models in mind, so the metrics are implemented to be differentiable if needed. While that is useful in its own regard that makes including new metrics a potential hassle because they may need to be rewritten from scratch. In terms of MT evaluation metrics, it supports BLEU, ChrF, TER, and also the more recent neural evaluation metric BERTScore (Zhang et al., 2020).

Similarly, `evaluate` by Hugging Face has a broad scope for different machine learning problems. At the start of this project, it already included many metrics to evaluate MT, including state-of-the-art ones, (BLEU, BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), BERTScore). How MATEO builds on both `sacrebleu` and `evaluate` will be discussed in more detail in the Section 3.

2.2 Web interfaces

Web interfaces for machine translation evaluation exist but they are often inactive, not maintained, or outdated. Asiya Online,[4] for instance, is an interface that was often used for educational purposes, but the underlying service is no longer available. Moreover, Asiya Online lacked support for recent, state-of-the-art metrics that I wanted to include. Tilde MT offers an interface for evaluating MT, but is restricted to

[2] https://github.com/BramVanroy/mateo-demo/

[3] MATEO is available at https://mateo.ivdnt.org/ (persistent identifier http://hdl.handle.net/10032/tm-a2-w6)

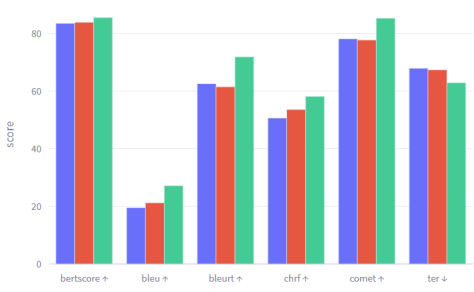
[4] https://asiya.cs.upc.edu/demo/asiya_online.php

calculating BLEU.[5] MT-ComparEval (Klejch et al., 2015) is an open-source evaluation interface that provides only BLEU, precision, recall, and F-scores. Although a demo is available, it seems to be restricted to previously uploaded data, with no option to intuitively analyze one's own data, which is fundamental to our proposal.

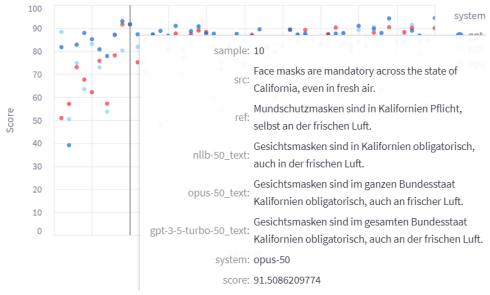
To provide users with the most intuitive visual experience, I draw inspiration from related, established research and computer-assisted translation (CAT) tools, in addition to the existing MT evaluation frameworks mentioned above. For instance, MateCat[6] is an online, open-source CAT tool that also offers attractive visual indicators to editors, similar in spirit to the CAT interface SCATE (Vandeghinste et al., 2019). In addition, the program CharCut(Lardilleux & Lepage, 2017) can give users a clear intuition of character-level differences between MT and reference translations by generating HTML web pages where these differences are highlighted.

3 MATEO

To build a web interface, first, a diverse array of both common and state-of-the-art evaluation metrics needed to be available. Initially, I endeavoured to make as many of them as possible available in the Python library `evaluate`. As part of the MATEO project and in the spirit of open source, I contributed multiple MT evaluation metrics to that library, most relevant here are TER, ChrF, and CharCut. Unfortunately, as the project progressed, it became clear via personal communication with the maintainers of `evaluate` that development of the library would be slowed down and new contributions that I may have in the future (new metrics, readable and deterministic metric signatures as in `sacrebleu`, statistical comparisons) were not guaranteed to be incorporated at a fast pace. In light of this unfortunate timing (early 2023), this forced me to make decisions on how to continue the project. The additional functionality that I required was therefore implemented in MATEO directly rather than via the `evaluate` library. In terms of code base, MATEO was built with extensibility in mind. As long as a metric is available in `evaluate` it is very straightforward for developers to add new metrics to it.



(a) Corpus level. Every group represents a metric, every bar a system.



(b) Sentence level. Every dot represents a sample, every color a system. An informative tooltip is shown on hover.

Figure 1: MATEO's interactive visualizations

To deliver a polished web interface, a Streamlit[7] multi-page web application was created. Under the hood it makes use of the current state of `evaluate`, which still provides a vast set of metrics, to access state-of-the-art neural metrics COMET, BERTScore, and BLEURT. The established lexical or character-based metrics TER, ChrF, and BLEU are included via a dependency on the aforementioned `sacrebleu`. In the interface, users can easily select metrics to use for the evaluation. Importantly, more advanced users also have the opportunity to change specific parameters that are available for each metric, for instance whether to ignore punctuation marks in calculating TER, lowercasing the texts in ChrF, or selecting a neural evaluation model in COMET. One can then upload text files to evaluate: one reference file that contains references translations, an optional file that contains source texts (needed when using the

[5] https://www.letsmt.eu/Bleu.aspx
[6] https://site.matecat.com/
[7] https://streamlit.io/

COMET metric), and up to four files containing different translations from different MT systems so that users can compare MT models to each other. The first of these MT files will be considered as the "baseline" translation. By clicking a button, the interface will calculate all the evaluation scores.

After calculation, a table with the results is given which includes the scores per system, which also includes confidence intervals and statistical significance scores. For every system it is indicated whether it differs significantly from the baseline system. This table can be downloaded as an Excel or tab-separated file. Specifically for researchers, the table can also be copy-and-pasted directly in their papers in LaTeX format. For metrics that support sentence-level scores (BERTScore, BLEURT, and COMET) a table is also provided for viewing and downloading that contains for every sample and every system an evaluation score, which is useful for fine-grained analysis. These results are shown to the users on both the corpus level and on the sentence level (Figure 1), and these visualizations can be downloaded as well.

Finally, the web interface offers an educational section on machine translation metrics, a translation component to generate baseline machine translations, and a page to visualize character and word differences between two texts, inspired by CharCut. This last feature enables an easy, visual analysis of machine translations.

4 Conclusion and project details

This paper introduces MATEO to the CLARIN audience. It is an open-access interface for evaluating MT, deliberately built for an audience with differing backgrounds, from advanced and technical users to novices and students. This project was kick-started with a Sponsorship 2021 grant from the European Association of Machine Translation (EAMT), and was subsequently awarded a substantial follow-up grant from the CLARIN.eu *Bridging Gaps* initiative. In writing the project proposal and brainstorming about the project, the author has received the welcome support of Lieve Macken and Arda Tezcan. The secured funding provided half-time employment for the author of this paper at Ghent University from July 2022 until June 2023.

References

Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022). Results of WMT22 Metrics Shared Task. *Proc. of the 7th WMT*, 46–68. https://aclanthology.org/2022.wmt-1.2

Klejch, O., Avramidis, E., Burchardt, A., & Popel, M. (2015). MT-ComparEval. *Prague Bulletin of Math. Ling.*, *104*(1), 63–74. https://doi.org/10.1515/pralin-2015-0014

Lardilleux, A., & Lepage, Y. (2017). CHARCUT. *Proc. of the 14th ISLT*, 146–153.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU. *Proc. of the 40th ACL*, 311–318. https://doi.org/10.3115/1073083.1073135

Popović, M. (2017). chrF++. *Proc. of the 2nd Conference on Machine Translation*, 612–618. https://doi.org/10.18653/v1/W17-4770

Post, M. (2018). A call for clarity in reporting BLEU scores. *Proc. of the Third Conference on Machine Translation: Research Papers*, 186–191. https://www.aclweb.org/anthology/W18-6319

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET. *Proc. of EMNLP 2020*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT. *Proc. of the 58th ACL*, 7881–7892. https://doi.org/10.18653/v1/2020.acl-main.704

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proc. of AMTA 2006*.

Vandeghinste, V., Vanallemeersch, T., Augustinus, L., Bulté, B., Van Eynde, F., ..., & Luyten, K. (2019). Improving the Translation Environment for Professional Translators. *Informatics*, *6*(2), 24. https://doi.org/10.3390/informatics6020024

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore. *Proc. of ICLR 2020*, 1–43. https://openreview.net/forum?id=SkeHuCVFDr

Domain-Specific Languages for Epigraphy: the Case of ItAnt

Federico Boschetti
CNR-ILC
URT Venezia, Italy
federico.boschetti@ilc.cnr.it

Luca Rigobianco
Dipartimento di Studi Umanistici
Università Ca' Foscari Venezia
Venezia, Italy
luca.rigobianco@unive.it

Valeria Quochi
CNR-ILC
Pisa, Italy
valeria.quochi@ilc.cnr.it

Abstract

ItAnt is a research project devoted to the languages and cultures of ancient Italy witnessed by a digital collection of inscriptions. This contribution illustrates how the definition of a Domain-Specific Language can support the activity of the epigraphists involved in the project by increasing the human readability of the encoded data without sacrificing the compliance to standard models and formats. Finally, an example of concrete use of the encoded texts within the CLARIN-IT DigItAnt platform will be briefly described.

1 Introduction

The project *Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models* (ItAnt hereafter) is an initiative funded by the Italian Ministry of University and Research and involves a consortium comprising the Ca' Foscari University of Venice, the University of Florence, and the Institute for Computational Linguistics "A. Zampolli" of the National Research Council of Italy[1]. This project aims at investigating the languages of Ancient Italy combining the methods of historical linguistics with digital technologies specifically designed to create a set of interrelated resources, particularly critical digital editions of inscriptions, lexica and bibliographies.

With the sole exception of Roman Latin, the languages of Ancient Italy (8th century BC-1st century AD) are fragmentary languages, that is to say dead languages attested through a highly restricted corpus of texts. Specifically, their evidence consists almost exclusively of epigraphic texts, which often present problems relating to the reading, segmentation into words, linguistic analysis, and interpretation. Therefore, one of the key challenges of the ItAnt project is to adapt the digital tools, practices, and methodologies of digital epigraphy and computational lexicography to the highly fragmentary nature of such a documentation.

The main objectives of the project are to create and interlink one another a digital archive of (critical editions of) inscriptions and a multilingual computational lexicon, as well as to encode project information using CIDOC CRM and its extensions, namely CRMtex[2], CRMinf [3], and FRBRoo/LRMoo[4]. With regard to the digital archive, the inscriptions are being encoded in XML according the XML-TEI/EpiDoc schema[5]. Furthermore, the edition of the inscriptions is enriched with standard metadata, thus allowing for an accurate description of each of them as both a linguistic and a material object.

2 The Digital Edition of the Inscriptions

As mentioned above, the project envisages the inscriptions being encoded according to the XML-TEI/EpiDoc schema. Such a schema is the result of an international effort aimed at customising the Text Encoding Initiative's standard for the representation of ancient documents according to the Leiden

[1] https://www.prin-italia-antica.unifi.it/
[2] https://www.cidoc-crm.org/crmtex/home-8
[3] https://www.cidoc-crm.org/crminf/
[4] https://cidoc-crm.org/frbroo/
[5] http: //www.stoa.org/epidoc/gl/latest/

Conventions. In particular, XML-TEI/EpiDoc provides mark-ups for the text (edition, apparatus, translation, commentary, bibliography) as well as the materiality and history of the object on which the text appears (repository, support, layout, hand, place and date of origin, provenance). Furthermore, thanks to the extensibility of XML and the versatility of XML-TEI/EpiDoc, ItAnt has proposed solutions for managing specific issues arising from the nature of the languages of Ancient Italy as fragmentary languages and their specific epigraphic features, with particular reference to word division, shape and reuse of the support, opisthography, line dimensions, description of linguistic elements, and description of languages and scripts (Murano et al., 2023).

With the goal of data integration, ItAnt makes use of widely used vocabularies and gazetteers, in particular The Art and Architecture Thesaurus provided by the The Getty Research Institute for the object type [6], material, and writing technique [7], the EAGLE vocabulary for the type of inscriptions (dedicatory, funerary, etc.)[8], and Pleiades and GeoNames for ancient and modern names respectively[9]. In addition, Trismegistos IDS are used, when available, to identify the text[10] and bibliographical records are also linked through a specific library built up by using Zotero[11].

3 How ItAntDSL Facilitates the Encoding

Encoding epigraphic contextual metadata and textual data in XML-TEI/EpiDoc is a complex, error-prone task. Indeed, XML-TEI is quite verbose (because element names, attributes and values must be written in full) and redundant (because opening and closing tags repeat the element names). The percentage of informative and structural contents is unbalanced. XML-TEI ensures data interchange among software applications and promotes machine actionability and interpretability, but human readability of an encoded document decreases rapidly as complexity increases.

In ItAnt linguistic, philological and prosopographical data are highly entangled. Each word is associated to its part of speech, conjectural integrations to textual gaps (lacunae) are recorded, and named entities are identified. These chunks of information often overlap: for instance a lacuna in a line of text may extend between the end of the third token and the beginning of the forth one, whereas a named entity defined by *praenomen* (partially conjectured), *gentilicium* and *patronymicus* may extend from the forth to the sixth token.

The problem of overlapping hierarchies in TEI is well-known and many solutions are available, both through manual encoding of stand-off annotations in XML (Spadini & Turska, 2019) and through alternative representations (e.g. in json), currently or planned to be convertible in XML-TEI (Neill & Schmidt, 2021). An experimental solution adopted in ItAnt for encoding part of the corpus, is based on a domain-driven approach, which involves the epigraphists to co-design a Domain-Specific Language (Parr, 2009), named ItAntDSL, to encode data and metadata. The aims of this approach are twofold: a) optimising the human readability of the encoded documents, both during manual encoding and for subsequent uses of the documents, and b) complying with the EpiDoc abstract model. As shown in Fig. 1, in fact, the encoding of contextual metadata (on the left) and textual data (on the right) is very compact. ItAntDSL is defined by a Context-Free Grammar (CFG) available on github[12]. The documents encoded in ItAntDSL are parsed by ANTLR (Parr, 2013), which first converts the Domain-Specific Language into XML with a proprietary schema (XML-ItAnt), based on the production rules of the CFG.

Then, a chain of xquery scripts and XSLT stylesheets transforms XML-Itant documents into XML-TEI/EpiDoc documents. The transformations are not limited to the translation of element names and to structural modifications, but extend to the integration of a) automatically generated IDs; b) default values omitted in ItAntDSL documents; c) expansion of complex structured data encoded in ItAntDSL

[6]The iDAI.thesauri provided by the Deutsches Archäologisches Institut (http://thesauri.dainst.org/de.html) is used as a supplement with regard to natural supports such as cliffs.

[7]https://www.getty.edu/research/tools/vocabularies/aat/

[8]https://www.eagle-network.eu/resources/vocabularies/typeins/

[9]https://pleiades.stoa.org/; https://www.geonames.org/

[10]https://www.trismegistos.org/tm/

[11]https://www.zotero.org/groups/2552746/

[12]https://github.com/CoPhi/itantdsl/

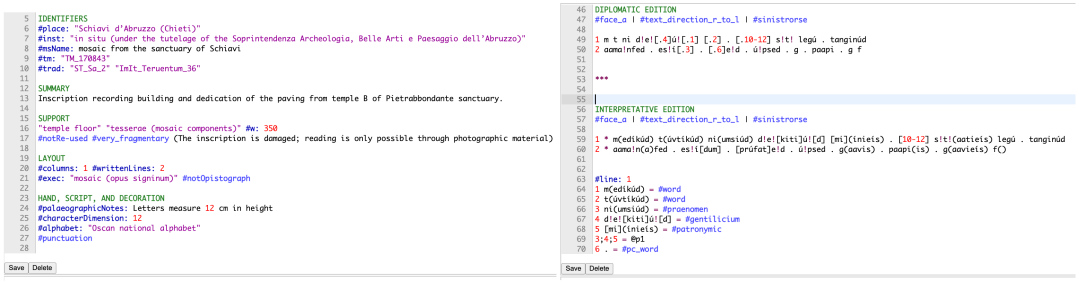


Figure 1: ItAntDSL

documents by reference (between quotation marks) and retrieved from the XML documents stored in an eXist-db. A sample of the final result is shown if Fig. 2.

```
<tei:div type="edition" subtype="interpretative" xml:space="preserve">
  <tei:div type="textpart" n="face_a" style="text-direction:r-to-l" rend="ductus:sinistrorse">
    <tei:ab>
      <tei:lb n="1" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1"/>
        <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">m</tei:supplied></tei:abbr><tei:ex>ediķúd</tei:ex></tei:expan>
        </tei:w>
        <tei:w xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_2">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">t</tei:supplied></tei:abbr><tei:ex>úvtíkúd</tei:ex></tei:expan>
        </tei:w>
        <tei:name type="praenomen" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_3" ref="#p1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">ni</tei:supplied></tei:abbr><tei:ex>umsiúd</tei:ex></tei:expan>
        </tei:name>
        <tei:name type="gentilicium" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_4" ref="#p1">
          <tei:unclear>de</tei:unclear>
          <tei:supplied reason="lost" evidence="previouseditor">kiti</tei:supplied>
          <tei:unclear>ú</tei:unclear><tei:supplied reason="lost" evidence="previouseditor">d</tei:supplied>
        </tei:name>
        <tei:name type="patronymic" xml:lang="osc-Ital-x-oscetr" xml:id="Osc_3_l_1_w_5" ref="#p1">
          <tei:expan><tei:abbr><tei:supplied reason="lost" evidence="previouseditor">mi</tei:supplied></tei:abbr><tei:ex>ínieís</tei:ex></tei:expan>
        </tei:name>
        <tei:pc unit="word">.</tei:pc>
        <!-- ... -->
```

Figure 2: XML-TEI/EpiDoc

4 ItAnt and CLARIN-IT: a Concrete Use-Case

ItAnt is developing a user-friendly web platform, DigItAnt, for creating, exploring and querying LOD-compliant lexica natively interlinked with critical editions of inscriptions, citations and bibliographic references, plus other external available salient resources. The DigItAnt editing application (*EpiLexO*) is meant to be useful especially for encoding lexical information of ancient languages and in assisting scholars in linking it to other relevant (re-)sources according to the semantic web principles. Particularly important and central to the platform is linking to the texts encoded in digital scholarly editions of relevant inscriptions in TEI-EpiDoc. The editions encoded with ItAntDLS and then converted to EpiDoc-XML as described above, are subsequently ingested by the platform, for linking to the lexicon and searching into the exploration platform (see Quochi, Bellandi, Mallia, et al. (2022) for reference). In details, the (historical linguist) user uploads one or more EpiDoc XML documents into the platform so that (s)he can then link the exact text loci to either existing or newly created lexical items (see Fig. 3 and Quochi, Bellandi, Khan, et al. (2022)).

Thanks to the EpiDoc-XML encoding, visualisation of contextual information as well as of the text according to the Leiden conventions is also possible, both in the Epilexo editor and in the exploration platform (*EpiLexO-search*, see an example in Fig. 4)

The platform is supported by CLARIN-IT and will become one of its services once the project is completed. The editing components of the platform are ready and in use within the project[13], while the

[13]The code is available at https://github.com/DigItAnt/Epilexo

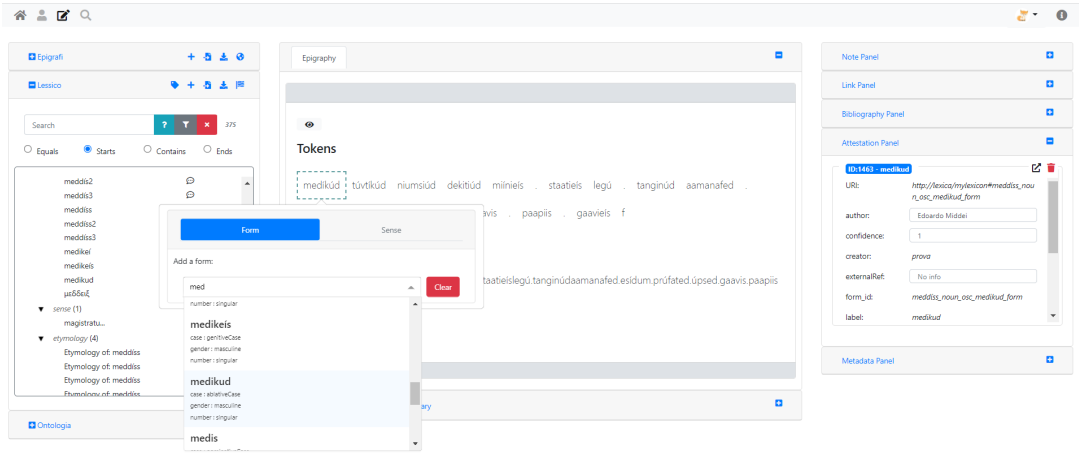


Figure 3: Linking texts loci to lexical forms in DigItAnt - EpiLexO

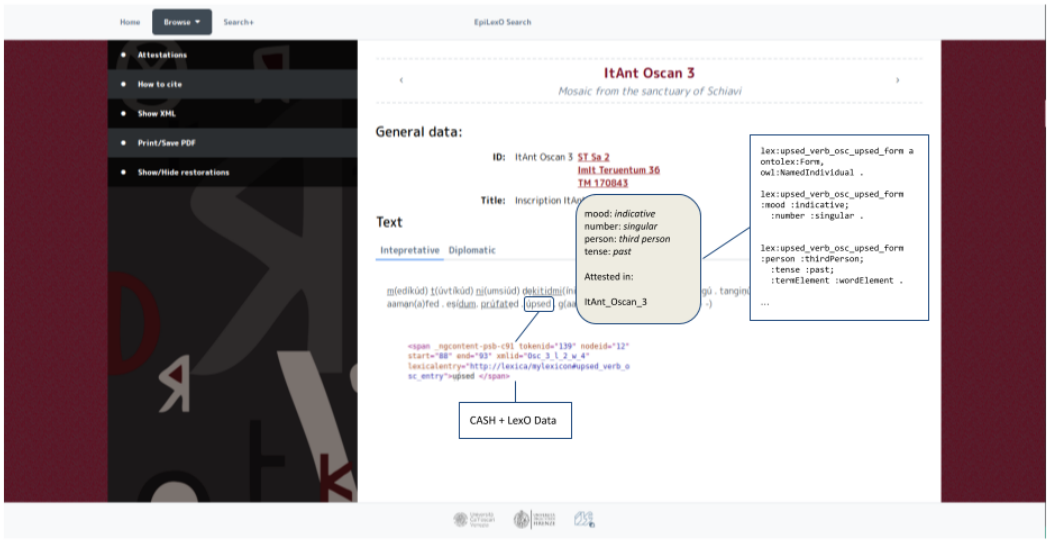


Figure 4: Linking texts loci to lexical forms in DigItAnt - EpiLexO

exploration platform is still in progress, continuously upgraded with new functionalities and improvements. An alpha version will be shown at the conference. Furthermore, as concerns the relations with CLARIN, all project resources, data and tools will be deposited into the ILC4CLARIN repository[14], in order to ensure their FAIRness, long-term preservation and maximal exploitability by the community[15].

Finally, the know-how related to the annotation of inscriptions through ItAntDSL will be shared through the Digital and Public Textual Scholarship Knowledge Centre[16] (DiPText-KC) of CLARIN.

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

[15]Currently, we deposited the first versions of ItAntDSL (http://hdl.handle.net/20.500.11752/ILC-1003) and EpiLexO (http://hdl.handle.net/20.500.11752/ILC-1004). At the end of the project all software, inscription corpora and ancient lexicons will be preserved, discoverable and consumable via CLARIN channels.

[16]https://diptext-kc.clarin-it.it/

Videotutorials and other initiatives, such as webinars and workshops, are planned towards the end of the project and after.

5 Conclusion

VeDPH, CNR-ILC and ILC4CLARIN in the last years are collaborating on DH projects related to many kinds of resources, such as collections of literary texts (Boschetti et al., 2021) and collections of epigraphic sources (Vagionakis et al., 2022). ItAnt provides a good opportunity to develop methods and tools to facilitate the encoding activities of the epigraphists, which must deal with complex entangled data. CLARIN provides not only the infrastructure to deposit the research data, but also the instruments to share new practices adequate to the domain of the epigraphic studies.

Acknowledgments

This work is supported by the Italian Ministry of the University and Research with the Italian National Strategic Research Grant PRIN 2017XJLE8J for the project: *Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models*. The project is also supported by and contributing to CLARIN-IT.

References

Boschetti, F., Del Grosso, A. M., & Spinazzè, L. (2021). La galassia musisque deoque: Storia e prospettive. In *Paulo maiora canamus - raccolta di studi per paolo mastandrea* (pp. 405–419). Edizioni CaFoscari. https://edizionicafoscari.unive.it/media/pdf/books/978-88-6969-558-2/978-88-6969-558-2-ch-26.pdf

Murano, F., Quochi, V., Del Grosso, A. M., Rigobianco, L., & Zinzi, M. (2023). Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process. *Journal on Computing and Cultural Heritage*, *16*, 1–14. https://doi.org/10.1145/3593431

Neill, I., & Schmidt, D. (2021). SPEEDy. A Practical Editor for Texts Annotated with Standoff Properties. *Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing*, *15*, 45.

Parr, T. (2009). Language implementation patterns: create your own domain-specific and general programming languages. *Language Implementation Patterns*, 1–380.

Parr, T. (2013). The definitive ANTLR 4 reference. *The Definitive ANTLR 4 Reference*, 1–326.

Quochi, V., Bellandi, A., Khan, F., Mallia, M., Murano, F., Piccini, S., Rigobianco, L., Tommasi, A., & Zavattari, C. (2022). From Inscriptions to Lexicon and Back: A Platform for Editing and Linking the Languages of Ancient Italy. *Proceedings of Second Workshop on Language Technologies for Historical and Ancient Languages LT4HALA 2022*, 59–67.

Quochi, V., Bellandi, A., Mallia, M., Tommasi, A., & Zavattari, C. (2022). Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO. *CLARIN Annual Conference Proceedings*, 39.

Spadini, E., & Turska, M. (2019). XML-TEI Stand-off Markup: One Step Beyond. *Digital Philology: A Journal of Medieval Cultures*, *8*(2), 225–239.

Vagionakis, I., Del Gratta, R., Boschetti, F., Baroni, P., Del Grosso, A. M., Mancinelli, T., & Monachini, M. (2022). 'Cretan Institutional Inscriptions' Meets CLARIN-IT. *CLARIN Annual Conference*, 139–150.

Finding Dutch multiword expressions

Jan Odijk
Utrecht University, the Netherlands
j.odijk@uu.nl

Martin Kroon
Utrecht University, the Netherlands
m.s.kroon@uu.nl

Tijmen Baarda
Utrecht University, the Netherlands
t.c.baarda@uu.nl

Ben Bonfil
Utrecht University, the Netherlands
b.bonfil@uu.nl

Sheean Spoel
Utrecht University, the Netherlands
s.j.j.spoel@uu.nl

Abstract

We present MWE-Finder, which enables a user to search for occurrences of multiword expressions (MWEs) in large Dutch text corpora. Components of many MWEs in Dutch can occur in multiple forms, need not be adjacent, and can occur in multiple orders (such MWEs are called *flexible*). Searching for occurrences of such flexible MWEs is difficult and cannot be done reliably with most search applications. What is needed is a search engine that takes into account the grammatical configuration of the MWE. MWE-Finder is therefore embedded in GrETEL, a treebank search application for Dutch. A user can enter an example of a MWE in a specific canonical form, after which the system searches for sentences in which the MWE occurs, using queries generated automatically from the canonical form. The MWE can also be selected from a list of more than 11k canonical forms for Dutch MWEs that MWE-Finder offers. We will show that MWE-Finder also offers facilities to find examples with unexpected modifiers or determiners on components of the MWE

1 Introduction

A multiword expression (MWE) is a word combination with linguistic properties that cannot be predicted from the properties of the individual words or the way they have been combined by the rules of grammar (Odijk, 2013). A word combination can, for example, have an unpredictable meaning (*de boeken neerleggen*, lit. 'to put down the books', meaning 'to declare oneself bankrupt'), an unpredictabe form (e.g. *ter plaatse* 'on location', with idiosyncratic use of *ter* and *e*-suffix on the noun), or it can have only limited usage (e.g. *met vriendelijke groet* 'kind regards', used as the closing of a letter). In a translation context, it can have an unpredictable translation (*dikke darm* lit. 'thick intestine', 'large intestine'), etc.

Many Dutch multiword expressions (MWEs) are flexible in the sense that their components can have different forms, can occur in different orders, or can have words that do not belong to the MWE between them. This flexible nature of such MWEs makes it difficult to reliably search for occurrences of such expressions in text corpora. Standard search engines such as Google do not enable the user to systematically search for different word forms of the same lemma. Search applications developed in the context of CLARIN such as OpenSoNaR (de Does et al., 2017; van de Camp et al., 2017) or Nederlab (Brugman et al., 2016) can do this, but it is difficult to formulate a query allowing different orders and interspersed irrelevant words, and the results of such a query will be unreliable. At best, one can find all instances but one will at the same time find many instances where all these words occur but not the MWE. One should be able to search for a flexible MWE in such a way that its grammatical structure is taken into account.

We present MWE-Finder, which is intended as a research tool for any linguist or lexicographer interested in research into multiword expressions, in particular *flexible* multiword expressions. MWE-Finder can take the grammatical structure of a MWE into account because it is embedded in a new version (version 5) of GrETEL,[1] an existing web application for searching Dutch treebanks developed in the context

[1] https://gretel5.hum.uu.nl

of CLARIN (Augustinus et al., 2012; Augustinus et al., 2017; Odijk et al., 2018). The distinguishing feature of GrETEL is its query-by-example feature. In its regular search mode, it leads the user through a number of steps to get from an example sentence to search results and analysis of the search results. MWE-Finder mimics this approach specifically for MWEs. We describe the relevant steps in section 2.

A second important feature of GrETEL is that one can upload one's own text corpus, which is then automatically parsed and made available as a treebank to search in. This feature is therefore automatically also available for MWE-Finder.

GrETEL is open source and its code is available on GitHub.[2] The part of the application that generates queries for MWEs and that performs the tree manipulation is available as a separate Python package, so that it may also be used to create scripts that search treebanks without using GrETEL.[3] We are in contact with researchers of the Institute for the Dutch Language (INT) to host GrETEL5 when it is completely finished at the recognized CLARIN Type B-center INT.[4]

2 MWE-Finder

The user goes through a number of steps to obtain the desired results: (1) example MWE; (2) treebank selection; (3) query results ; (4) analysis of the query results. We describe each of these steps here.

2.1 Example MWE

MWE-Finder enables a user to search for occurrences of a MWE in a treebank based on an example MWE. The example MWE must be in a specific canonical form. For single words the canonical form is its lemma. However, for reasons that will be described in the full paper, in many verbal MWEs one cannot use just the lemma for the head of the MWE. Instead, we require that a verbal MWE is an infinitival complement to the future auxiliary verb *zullen* 'will'. A concrete example is (1) in which the indefinite pronoun *iemand* in the canonical form means that any phrase can occur here:

(1) iemand zal de dans ontspringen
someone will the dance spring
'someone will have a lucky escape'

It is assumed that the head of the MWE can be inflected, modified and determined, but that other parts of the MWE cannot. Of course, there are many exceptions to this, and these are indicated by means of annotations. There are annotations to mark (un)modifiability and (un)inflectability of MWE components, not being a component of the MWE, specific limited types of variation, and for variables parts of the MWE, etc., as will be explained in detail in the full paper. With the canonical form of the MWE the user implicitly formulates a hypothesis about the properties of this MWE.

MWE-Finder offers the user a large list of MWEs in canonical form to select from. This list was derived from the DUtch CAnonicalised Multiword Expressions (DUCAME) resource.[5]

2.2 Treebank selection

The user selects the treebank or treebanks that the MWE should be searched in. As a concrete example, one could choose the treebank MEDIARGUS, which contains texts from Belgian newspapers (more than 103 million sentences).

2.3 Query results

MWE-Finder automatically generates three queries from the MWE example in canonical form to search for occurrences of this MWE in a treebank. These are the *major lemma query*, the *near-miss query*, and the *MWE query*.[6] The query generation process has been described in detail in (Odijk et al., to appear).

[2] https://github.com/UUDigitalHumanitieslab/gretel
[3] https://github.com/UUDigitalHumanitieslab/mwe-query
[4] https://www.ivdnt.org/.
[5] See (Odijk et al., to appear) and https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP.
[6] Note that MWE-Finder can identify potential occurrences of a MWE in a treebank. It cannot determine for an expression that is ambiguous between a literal and an idiomatic reading which of these alternative readings is applicable in a specific sentence.

Once selected, the application switches to the *Results* view where query results are displayed as they arrive from the server. In that view, the user can also switch between the different queries for the chosen MWE or choose to exclude results of finer-grained queries.

The major lemma query searches for sentences in which at least the lemmas of the major words of the MWE occur (in any grammatical configuration). Major words are the content words if there are at least two in the MWE, and content and function words if there is at most one content word in the MWE. The query yields a superset of the results of both other queries. This query is applied to the full treebank, making use of indexes on the treebank to speed up the process. The major lemma query yields a list of syntactic structures, and can be used to identify the MWE in a grammatical configuration that was not expected at all, to retrieve occurrences of the MWE in sentences that the Alpino parser used in GrETEL parsed incorrectly, or to retrieve occurrences of the MWE for which MWE-Finder did not generate the correct other two queries on the basis of the canonical form. The syntactic structures in the output of the major lemma query are modified in ways described in (Odijk et al., to appear). The near-miss query and the MWE query are applied to the modified output of the major lemma query.

The near-miss query searches for sentences in which the lemmas of the major words of the MWE occur in the grammatical configuration derived from the canonical form. It can find potential examples of the MWE that deviate from the canonical form provided by showing differences in forms, arguments, modification and determination. It yields a superset of the MWE query results and can be used to fine-tune the hypothesis on the MWE as encoded in the canonical form supplied by the user.

The MWE query finds sentences in which the MWE occurs. This query takes into account the hypothesis on the MWE implied by the canonical form and its annotations supplied by the user.

For the canonical form (1), applied to the MEDIARGUS treebank, the results are as follows:[7] The major lemma query yields 1309 results. The near-miss query yields 1271 results. The MWE query yields 1158 results

2.4 Analysis

Finally, there is the analysis step, which is identical to the one in GrETEL.

For a MWE, one would additionally like to analyse the result set in ways that cannot be done by GrETEL's standard analysis component. We are working on adding a special analysis step for MWEs, in which the system gathers statistics on the components of the MWE, the arguments of the MWE (their grammatical relations and syntactic categories, and their heads), the argument frames[8] that occur with the MWE, and about modifiers and determiners for the MWE as a whole and for each of its components. It does this for the results of the MWE query, for the results of the near-miss query, and for the difference between the near-miss query and the MWE query. We have an initial version available but at the time of writing it has not been integrated yet in the actual application.

But even without this dedicated analysis component MWE-Finder enables the user to quickly analyse the search results. In the results of the near-miss query one can exclude the results of the MWE query, leaving only 113 results for manual inspection. Even a cursory look shows that different determiners than *de* occur with *dans* (in particular *die*), that the determiner can be absent (but apparently only in headlines), and that the word *dans* can be modified by adjectives (e.g. *gerechtelijke* 'judicial', *fiscale* 'fiscal', *fatale* 'fatal') and by PPs (e.g. *van het faillissement* 'of the bankruptcy'). Several other results are due to wrong parses.

In the results of the major lemma query one can exclude the results of the near-miss query, which leaves 38 utterances for manual inspection. Most of these involve wrong parses, some examples involve the variant *aan de dans ontspringen*.

These result convince us to revise our hypothesis on the expression *de dans ontspringen* as implicit in the canonical form. Better canonical forms for this expression are probably *iemand zal dd:[de] *dans*

[7] The query names are links to the actual queries. All queries last retrieved 2023-03-20.

[8] With *argument frame* we mean a list of (extended relation, syntactic category) pairs for the arguments that the MWE occurs with, where an extended relation is a sequence of grammatical relations. For example, in *Marie brak Piets hart.* lit. 'Marie broke Piet's heart.', the argument frame is [(su, NP), (obj1/det, NP)], i.e., it combines with two arguments, a subject NP and a NP functioning as the determiner of the direct object.

ontspringen or *iemand zal 0de *dans ontspringen*, where *dd:[de]* means that *de* can be replaced by any definite determiner, **dans* means that the word *dans* is modifiable, and *0de* means that the word *de* is not part of the MWE. Furthermore, we found a variant of this MWE, with canonical form *iemand zal aan 0de *dans ontspringen*.

3 Limitations

MWE-Finder is fully dependent on the syntactic structures generated by the Alpino parser. If Alpino cannot parse a sentence correctly, MWE-Finder will not be able to identify any MWE in it. This is one of the reasons why MWE-Finder includes the major lemma query: this query will find sentences in which the MWE occurs even if Alpino cannot parse it correctly, so a researcher still has data to work with.[9] However, this query will also find many sentences in which the MWE does not occur, so it will require more manual work by the researcher. We aim to reduce the amount of manual work required by providing statistics on the results and the results minus the results of the other two queries in the dedicated MWE Analysis step. In particular, it will provide statistics on the grammatical relation between the lemmas of the major words. However, at the time of writing this has not been integrated in the online version yet.

4 Conclusions

MWE-Finder makes it possible to reliably and quickly search for occurrences of a MWE despite its flexible nature. We submit that MWE-Finder is a useful research instrument for linguistic and lexicological research into MWEs, and can form an exemplary research instrument in the CLARIN research infrastructure.

References

Augustinus, L., Vandeghinste, V., & Eynde, F. V. (2012). Example-based treebank querying. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)*. European Language Resources Association (ELRA).

Augustinus, L., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2017). GrETEL: A tool for example-based treebank mining [DOI: http://dx.doi.org/10.5334/bbi.22. License: CC-BY 4.0]. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 269–280). Ubiquity.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Sang, E. T. K., & van den Bosch, A. (2016). Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. European Language Resources Association (ELRA).

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab [DOI: http://dx.doi.org/10.5334/bbi.20. License: CC-BY 4.0]. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 245–257). Ubiquity.

Odijk, J. (2013). Identification and lexical representation of multiword expressions. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch. Results by the STEVIN-programme* (pp. 201–217). Springer. http://link.springer.com/content/pdf/10.1007

Odijk, J., Kroon, M., Baarda, T., Bonfil, B., & Spoel, S. (to appear). MWE-finder: Querying for multiword expressions in large Dutch text corpora. In V. Giouli & V. B. Mititelu (Eds.), *Multiword expressions in lexical resources. linguistic, lexicographic and computational perspectives*. Language Science Press.

Odijk, J., van der Klis, M., & Spoel, S. (2018). Extensions to the GrETEL treebank query application [http://aclweb.org/anthology/W/W17/W17-7608.pdf]. In *Proceedings of the 16th international workshop on treebanks and linguistic theories (tlt16)* (pp. 46–55).

[9] Assuming Alpino can at least lemmatize all major words correctly.

van de Camp, M., Reynaert, M., & Oostdijk, N. (2017). WhiteLab 2.0: A web interface for corpus exploitation [DOI: http://dx.doi.org/10.5334/bbi.19. License: CC-BY 4.0]. In J. Odijk & A. van Hessen (Eds.), *Clarin in the low countries* (pp. 231–243). Ubiquity.

Automatic Anonymization of Human Faces in Images of Authentic Social Interaction: A web application

André Frank Krause
University of Duisburg-Essen
andre.krause@uni-due.de

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

Karola Pitsch
University of Duisburg-Essen
karola.pitsch@uni-due.de

Abstract

To provide easy access to anonymization tools, an open-access and open-source web application is described that employs state-of-the-art machine learning models for automatic face and head-region anonymization. Besides anonymization, the application can provide useful data for multimodal interaction research, like body postures and face locations.

1 Introduction: The Need for Anonymizing Data of Authentic Social Interaction

Empirical research on human social interaction requires the researcher to respect ethical and legal issues of data collection and management (Roth, Von Unger, et al., 2018). In particular, when creating video corpora of authentic multimodal communication, researchers need to be concerned with the protection of personal data both on the auditory and visual level (Rubinstein & Hartzog, 2016).

The European General Data Protection Regulation (GDPR, in Germany also known as the ”Datenschutz-Grundverordnung” - DSGVO) defines legal regulations that apply to all organizations in the European Union, including scientific research facilities. Wiewiorówski, 2020 and Bäcker and Golla, 2020 discuss the impact of the GDPR on scientific and commercial research. For example, Article 89 of the GDPR requires that, if processing methods are available to avoid the identification of data subjects, those measures must be applied, except when those methods seriously impair the achievement of the specific purposes. In Watteler and Ebel, 2019, different levels of anonymization and the important difference between absolute and de-facto anonymization are elaborated. The latter aspect is important for archiving and data reuse.

While there is a long tradition of anonymizing / pseudonymizing transcripts of the spoken word (Kretzer, 2013) and tools exist that help the researcher do so practically (e.g. Nicolai et al., 2021), there is little information and technical support on how to best do this for video data. On the textual level, relevant personal data to be protected concern the names of people, places, streets, federal states, and institutions, the professional and educational background of participants, as well as time information and more indirect context information (e.g. Kretzer, 2013). These pieces of critical information are usually manipulated both in text and auditory files by overwriting them with an ”xxxx”, overlaying the spoken word with a ’beep’ or the like. Anonymization of voice constitutes another urgent topic exhibiting conflicting objectives (privacy vs. utility for linguistic and interactional research, see e.g. (Srivastava et al., 2022)). In video recordings of social interaction, the visual level also needs consideration. For example, people can potentially be deanonymized using facial recognition or highly specific biometric identifiers like the iris pattern (Daugman, 2006).

All current image and video processing tools offer a range of filters that can be overlaid over the original video or image. On the one hand, applying these filters to the entire image or video helps to save time, but also often makes relevant setting information invisible. On the other hand, manually selecting only the face region of participants and applying a filter, is highly time-consuming work and requires some basic image or video editing skills. This seems to be a common hurdle for correctly anonymizing data from authentic social interaction.

(a) conf=0.95, nms=0.75 (b) conf=0.75, nms=0.75 (c) conf=0.05, nms=0.75 (d) conf=0.05, nms=0.25

Figure 1: Effects of the two main face detection parameters. Lowering the confidence threshold (conf) increases the number of detected faces (compare (a, b and c)). Multiple detections of the same face can be reduced using non-maximum suppression (nms, compare (c) with (d)).

In this paper, we address this gap and report progress on developing an open-source, sustainable web application that combines current face- and body-posture detection algorithms to automatically anonymize people in images and video frames. The application is part of a larger toolkit for anonymizing video recordings of social interaction (Krause et al., 2023a, 2023b) and will be freely available as open-source software here: https://git.uni-due.de/mumocorp-open-access/anonymization. It has been developed within the data-reuse project "MuMoCorp" [1] which focuses explicitly on preparing existing data on human-robot interaction to be accessible for other researchers. It will be, in part, integrated and published within an institutional repository and stored in a long-term repository, namely the IDS Repository[2], which is a member of CLARIN.

2 Approaches for Detecting Faces in Data from Authentic Human Social Interaction

In recent times, attempts have been made to use novel techniques of motion tracking and feature detection to advance research on multimodal communication and social interaction (Pfänder & Couper-Kuhlen, 2019; Pitsch et al., 2014). In this paper, we suggest exploring the potential of current feature detection and machine learning (ML) techniques for the issue of anonymizing people's data in video recordings and/or still images of authentic social interaction. Given the current technical capabilities, the question arises whether only a participant's face or the entire body should be anonymized and whether the relevant parts should be made unidentifiable using classic methods (e.g., blurring) or whether they should be replaced with a deep-fake. This latter option is suggested by recent state-of-the-art frameworks like Deep Privacy 2 (Hukkelås & Lindseth, 2022). At first sight, deep-fakes maximize privacy. But such an approach reduces, removes, or - most problematic for analytic purposes - modifies social cues. For example, the deep-fake approach used in the Deep Privacy 2 framework can lead to different gaze directions, modified hand and body postures, and different gender, age and nationality. Hence, there is difficulty achieving a good balance between an acceptable level of anonymization and the potential loss of social cues that are important for interaction research.

Our approach is based on a combination of two different ML methods to localize the head region of a person visible in an image. The detected region is then overlaid with a specific filter (at the current stage: a blur filter) to anonymize the person. The first method employs deep-learning based face detectors. The user can choose between RetinaFace (Deng et al., 2020) and the Dual Shot Face Detector (Li et al., 2019). RetinaFace implements robust and fast single-shot face detection. The model works reliably for fully visible faces, but might fail for faces that are either partially occluded or only partially visible from diagonally behind. The Dual Shot Face Detector (DSFD) was optimized for challenging face detection situations, including bad lighting conditions, reflections, unusual makeup, blurry faces, and unusual face orientations. DSFD provides excellent detection rates, but still occasionally misses detecting a face. To further reduce the chance of non-anonymized faces (false negatives), face detection is combined with the second method: human pose estimation using YOLO7 (Wang et al., 2022). Pose estimation provides several human-body key-points that can be used to infer the position and size of the head region.

[1] https://www.uni-due.de/kowi/mukom/mumocorp
[2] http://repos.ids-mannheim.de/

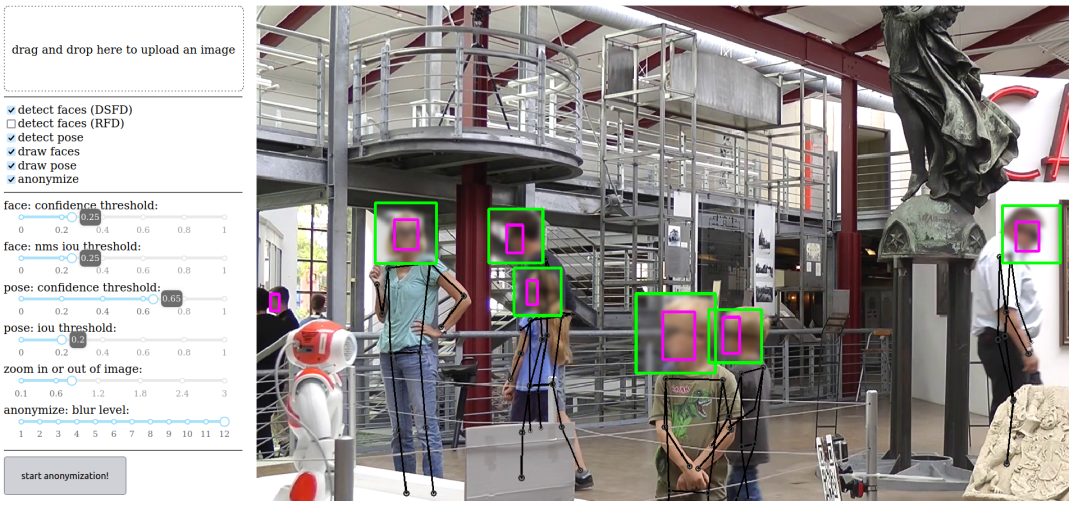


Figure 2: Screenshot of the web application. Several sliders allow the adjustment of detector parameters. The example image shows the result of head-region anonymization using a blurring filter. Magenta-colored boxes show detected faces, green boxes show the head region estimated using key points from pose tracking (black dots and lines).

2.1 Detection Parameters

The sensitivity of the face- and pose-detection methods can be adjusted and fine-tuned for the specific task requirements. The two main parameters (see also Fig. 1) offered by the frameworks are:

- Confidence threshold: ML-based methods often provide a confidence value for detected objects. Reducing the confidence threshold might include more detections, thereby increasing the number of detected faces (true positive rate), but typically will also increase the number of spurious detections, i.e., the false positive rate. For people anonymization, where the cost of a missed face detection (false negative) is much higher than a false-positive detection, the threshold parameter should be set as low as possible.

- Non-maximum suppression (NMS): Deep-learning based detectors may generate multiple, slightly different bounding boxes of varying confidence for the same object. NMS eliminates redundant bounding boxes and tries to select the optimal target boundary box (Gong et al., 2021).

2.2 Anonymization Efficiency

The anonymization efficiency of three different, state-of-the-art face- and pose-detection methods has been evaluated in a separate, more technical companion paper (Krause et al., 2023a). The detectors were tested on a challenging dataset with hand-selected frames from videos of a large multimodal corpus of human-robot interaction 'in the wild' (Pitsch, 2020). The dataset contains faces and heads of people in difficult-to-detect situations, with faces or people being partially occluded or overlapping, in unusual poses, motion blurred, or under bad lighting conditions. Some of the faces are barely visible from behind. Results show that the RetinaFace detector is slightly outperforms the DSFD detector if recall (maximizing anonymization) is favored over precision (minimizing false-positive face detections). Pose detector based head-region estimation yields an even better recall, but at the cost of comparatively low precision.

2.3 Web application: Features and Interface

The web application currently has two main features: 1. The anonymization of individual images (see fig. 2) and 2. The interactive adjustment and fine-tuning of detector parameters (see section 2.1). For example, a typical workflow for anonymizing a video would consist of several steps: 1. Extracting a few

typical frames from the video; 2. Uploading a frame to the web application and adjusting parameters; 3. Running the video anonymization tool. The workflow should also include a manual checking stage to annotate missed face detections, see (Krause et al., 2023a).

2.4 Web application: Implementation Details

The web app has been implemented using the programming language Python, which is currently gaining popularity as a web-development tool (Saabith et al., 2019). For face-detection, performance-optimized implementations of DSFD and Retinaface by Håkon Hukkelås are used (Hukkelås, n.d.). YOLO7 based pose estimation uses code from (Yiu, 2023). The browser-based user interface (HTML, CSS and JavaScript) is automatically generated by the Python framework Dash (Dabbas, 2021) and is "responsive", i.e., follows responsive design rules to ensure that a website is properly displayed on devices with different screen sizes. The Dash-based web app is then served by a Python-based, multithreaded WSGI web server called Gunicorn (Chesneau et al., n.d.). To improve security, a front-facing Apache web server is used, configured as a reverse proxy for Gunicorn.

3 Outlook

The web application must be optimized to improve multi-user performance. The architecture of the Dash framework is, by default, stateless to be able to scale and serve hundreds or thousands of users. This stateless approach requires that the comparatively large ML models be loaded from disk each time an anonymization is performed. This is neither performant nor memory-efficient. Especially if the models are executed on a GPU, memory could be exhausted quickly by only a few concurrent users. A future version of the tool should solve this problem using, e.g., a microservice architecture. Smaller improvements for the application may be the addition of additional anonymization filters and an option to adjust the filter strength based on the face area.

Another option could be the addition of more complex anonymization filters that may, e.g., use less blurring for preserving social cues but disturb automatic face recognition by, e.g., applying a defined amount of random face morphing, (Ferrara et al., 2022). Such filters should be carefully evaluated regarding their effects on privacy, both with respect to human and machine-based face identification capabilities.

Further tools will be implemented to not only process and anonymize images but also videos. Resulting body-posture trajectories, combined with 3D facial landmark localization (Khabarlak & Koriashkina, 2021) and gaze direction estimation (Ablavatski et al., 2020), will yield valuable data for in-depth multimodal interaction analysis and visualization.

Acknowledgments

This project was financed by the Volkswagenstiftung (grant number 90886, PI: Karola Pitsch).

References

Ablavatski, A., Vakunov, A., Grishchenko, I., Raveendran, K., & Zhdanovich, M. (2020). Real-time pupil tracking from monocular video for digital puppetry. *arXiv preprint arXiv:2006.11341*.

Bäcker, M., & Golla, S. (2020). *Handreichung datenschutz. 2. vollständig überarbeitete auflage* (tech. rep.). RatSWD Output. https://doi.org/10.17620/02671.50

Chesneau, B., Kapustin, K., Leeds, R., Peksağ, B., Madden, J., & Randall, B. (n.d.). Gunicorn - Python WSGI HTTP Server for UNIX — gunicorn.org [https://gunicorn.org/, Accessed 11-Apr-2023].

Dabbas, E. (2021). *Interactive dashboards and data apps with plotly and dash: Harness the power of a fully fledged frontend web framework in python–no javascript required*. Packt Publishing Ltd.

Daugman, J. (2006). Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proceedings of the IEEE*, *94*(11), 1927–1935.

Deng, J., Guo, J., Ververas, E., Kotsia, I., & Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5203–5212.

Ferrara, M., Franco, A., Maltoni, D., & Busch, C. (2022). Morphing attack potential. *2022 International Workshop on Biometrics and Forensics (IWBF)*, 1–6.

Gong, M., Wang, D., Zhao, X., Guo, H., Luo, D., & Song, M. (2021). A review of non-maximum suppression algorithms for deep learning target detection. *Seventh Symposium on Novel Photoelectronic Detection Technology and Applications*, *11763*, 821–828.

Hukkelås, H. (n.d.). Face-detection package — pypi.org [https://pypi.org/project/face-detection/, Accessed 12-Apr-2023].

Hukkelås, H., & Lindseth, F. (2022). Deepprivacy2: Towards realistic full-body anonymization. *arXiv preprint arXiv:2211.09454*.

Khabarlak, K., & Koriashkina, L. (2021). Fast facial landmark detection and applications: A survey. *arXiv preprint arXiv:2101.10808*.

Krause, A. F., Ferger, A., & Pitsch, K. (2023a). Anonymization of persons in videos of authentic social interaction: Machine learning model selection and parameter optimization. *10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023)*.

Krause, A. F., Ferger, A., & Pitsch, K. (2023b). Detecting and tracking persons in video recordings of authentic social interaction: Analysis and anonymization. [Computational and Quantitative Approaches to Multimodal Video Analysis - CAMVA 2023].

Kretzer, S. (2013). Arbeitspapier zur konzeptentwicklung der anonymisierungs-/pseudonymisierung in qualiservice.

Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., & Huang, F. (2019). Dsfd: Dual shot face detector. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5060–5069.

Nicolai, T., Mozygemba, K., & Hollstein, B. (2021). Qualianon–qualiservice tool für anonymizing text data (version 1.0. 1). *Retrieved from Software available at: https://github. com/pangaea-data-publisher/qualianon*.

Pfänder, S., & Couper-Kuhlen, E. (2019). Turn-sharing revisited: An exploration of simultaneous speech in interactions between couples. *Journal of Pragmatics*, *147*, 22–48.

Pitsch, K. (2020). Answering a robot's questions: Participation dynamics of adult-child-groups in encounters with a museum guide robot. *Reseaux*, *220221*(2), 113–150.

Pitsch, K., Vollmer, A.-L., Rohlfing, K. J., Fritsch, J., & Wrede, B. (2014). Tutoring in adult-child interaction: On the loop of the tutor's action modification and the recipient's gaze. *Interaction Studies*, *15*(1), 55–98.

Roth, W.-M., Von Unger, H., et al. (2018). Current perspectives on research ethics in qualitative research. *Forum qualitative sozialforschung/forum: Qualitative social research*, *19*(3), 1–12.

Rubinstein, I. S., & Hartzog, W. (2016). Anonymization and risk. *Wash. L. Rev.*, *91*, 703.

Saabith, A., Fareez, M., & Vinothraj, T. (2019). Python current trend applications-an overview. *International Journal of Advance Engineering and Research Development*, *6*(10).

Srivastava, B. M. L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., & Yamagishi, J. (2022). Privacy and utility of x-vector based speaker anonymization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2383–2395.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*.

Watteler, O., & Ebel, T. (2019). Datenschutz im forschungsdatenmanagement. *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten*, 57–80.

Wiewiorówski, W. (2020). A preliminary opinion on data protection and scientific research. *Brussels, Belgium: European Data Protection Supervisor*.

Yiu, W. K. (2023). Implementation of paper - YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [Online; accessed 12-Apr-2023]. https://github.com/WongKinYiu/yolov7

Building and consolidating a FAIR-compliant ecosystem of infrastructures

Noah Bubenhofer
Linguistic Research Infrastructure,
University of Zurich,
Switzerland
noah.bubenhofer@ds.uzh.ch

Andrea Malits
Language Repository of Switzerland,
Zurich University Library
Switzerland
andrea.malits@ub.uzh.ch

Stefanie Strebel
Language Repository of Switzerland,
Zurich University Library
Switzerland
stefanie.strebel@ub.uzh.ch

Johannes Gräen
Linguistic Research Infrastructure,
University of Zurich,
Switzerland
johannes.graen@linguistik.uzh.ch

Stefan Buerli
FORS,
University of Lausanne
Switzerland
stefan.buerli@fors.unil.ch

Cristina Grisot
CLARIN-CH Coordination Office,
University of Zurich
Switzerland
cristina.grisot@uzh.ch

Abstract

This presentation aims at showcasing the Swiss ecosystem of research infrastructures necessary for dealing with language resources, which is organized under the umbrella of Swiss node of CLARIN, known as CLARIN-CH. Since 2018, a consortium of partners has been working on building and consolidating a national ecosystem of infrastructures, which covers the whole linguistic data lifecycle from data generating, processing and analyzing to data sharing and archiving. This ecosystem includes : (i) the Linguistic Research Infrastructure which is a national technology platform hosted at the University of Zurich, (ii) the national repository for publishing and archiving linguistic data (iii) a database of Swiss media texts and a corpus platform for hosting of and searching in large text and audio/video corpora. The CLARIN-CH Coordination Office supports these infrastructures in reaching their mission and ensures communication and collaboration among the partners of ecosystem of research infrastructures, the members of the scientific community, the national bodies of funding and research politics, as well as with CLARIN ERIC.

1 Introduction

In line with the world-wide cultural change that took place in science in the last ten years to-wards Open Science, Switzerland has followed the path and implemented a national Open Research Data (ORD) programme to support the upgrading of existing infrastructures and the creation of new ones. Since January 2023, the Swiss ecosystem of research infrastructures (RIs) for linguistic, consisting of the national Linguistic Research Infrastructure, the repository for publishing and archiving linguistic data, a database of Swiss media texts and a Linguistic Corpus Platform, is the subject of a project carried out within the national ORD programme: the UpLORD project. The ultimate goal of this project is to

enhance the current ecosystem of infrastructures in order to facilitate ORD practices, as it is data shown in Figure 1.

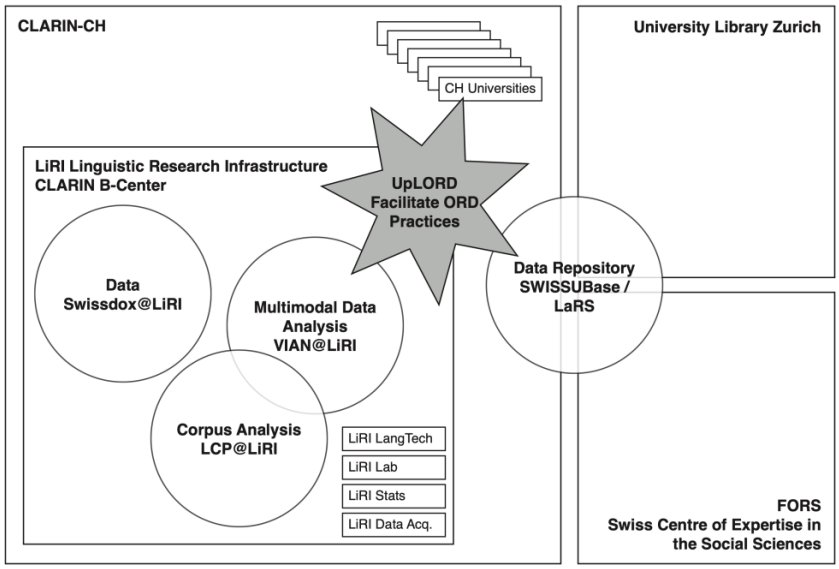


Figure 1 CLARIN-CH ecosystem of infrastructures and its upgrading in the UpLORD project

Through this project, the national ecosystem is being upgraded to answer, at the same time, to the needs of the target scientific communities – CLARIN-CH and European CLARIN – and to Open Science requirements. and the Linguistic Research Infrastructure and the data repository provide services to deal with each aspect of the life cycle of research data: *planning research >> collecting data >> processing and analysing data >> publishing and sharing data >> preserving data >> reusing data* (from the UK Data Service, cited by Mattern 2022: 61). The mission of this national ecosystem is to provide the Swiss scientific communities using linguistic data the services and the infrastructure necessary for their data to adhere to FAIR principles, and therefore to adopt sustainable practices that support sustainable linguistic research data, which in turn, will lead to replicable and sustainable research results.

2 Description of the CLARIN-CH ecosystem of infrastructures

2.1 The Linguistic Research Infrastructure LiRI

LiRI is conceived as a national platform which provides support by experts for researchers, students but also customers outside academia in the domains of data acquisition, processing, and analysis in linguistics and its subdisciplines. It includes groups for **language technology**, which proposes service in the area of Computational and Corpus Linguistics, such as collecting and processing language data, development and maintenance of purpose-built applications, long-term archival of research data, tailored workshops, consulting, and coaching, **a state-of-the-art lab** equipped with a bundle of data acquisition units, mainly for use in phonetics, psycholinguistics and neurolinguistics, and a group for **statistics and statistical consulting**. As such, LiRI proposes a comprehensive suite of services that covers every stage of a research project's life cycle, from the initial planning and experiment design to the acquisition and processing of data, as well as language technology services, and statistical consulting In 2019, LiRI was selected for the national Roadmap for Research Infrastructures and it is Switzerland's technical center. In the perspective of becoming a CLARIN B-center, an ecosystem of infrastructure is being built that promotes research according to ORD principles covering data contribution, data processing and data sharing and archiving.

2.2 Swissdox@LiRI

Swissdox@LiRI is the largest database of Swiss media texts and it is hosted at LiRI. It allows the distribution and the use for academic purposes of journalistic content despite its regular restrictive copyright conditions. The database contains a daily growing collection of at the moment 24 million media articles of more than 250 media titles. The data is provided in CSV format which allows further automatic processing, e.g. natural language processing. The contract with SMD Swiss Media Database allows free use of the data within an academic project, but only derivates of the data may be redistributed and archived. In order to allow reproducibility of analyses based on Swissdox@LiRI data and as much compatibility with the FAIR principles as possible, the query the user defined to get the data set is provided in JSON format and can be published and archived. This allows researchers whose institutions have access to Swissdox@LiRI to reproduce the data basis. Therefore Swissdox@LiRI is a key example of combining ORD principles with copyright constraints.

2.3 The Linguistic Corpus Platform LCP

The LCP is meant to be a platform for the hosting of complex linguistic annotated data allowing refined search and quantitative analyses. Infrastructures that allow hosting of corpora for online searching and querying the data are a key element of ORD. The LCP, which is planned to be launched in autumn 2023, will not only simplify work with corpora by offering even complex queries via an easy-to-use interface, but also facilitate the reproducibility and reusability of the analyses by other researchers. Queries can be named, stored, validated, visualised and shared. In the first release, support for most features of CQP's CQL query language is provided, with support for other query languages (such as ANNIS's AQL) coming in follow-up releases. Simple string searching and regular expression searchers are also possible. The software dependencies are as follows: Python 3 for backend, aiohttp for server, Vue.js for frontend, PostgreSQL for database, lark and cquery for Query parsing, and axios for HTTP requests. Import and export functions accept and offer standard data formats (XML-TEI, CoNNL) and all conceivable forms of annotation can be added as any number of layers. This fosters the reusability of these corpus data, e.g. by adding new annotation layers. The LCP also integrates VIAN-DH@LiRI, a software application that allows multi-modal communication and other linguistic domains, such as: (i) facilitates video processing and scaling over large data sets by using automation, (ii) implements automatic AI tools for speech, image recognition and NLP, (iii) proposes standardized annotation and methodology for quantitative multimodal analysis, (iv) develops query options for multimodal corpora.

2.4 The national repository SWISSUbase and the Language Repository of Switzerland LaRS

SWISSUbase & LaRS is a national and cross-disciplinary free data repository. It allows users to explore a cross-disciplinary catalogue, get data, publish studies and deposit data. Users find on the SWISSUbase website general and domain-specific user guides, guidelines and other helpful materials for depositing their data on SWISSUbase, and may benefit of personalised support from the various Data Service Units. To achieve a shared vision of an ORD landscape for linguistic research data in Switzerland and to design and develop federated services, LiRI and SWISSUbase & LaRS have established a close collaboration to offer the final step in the life cycle of research: publishing and archiving of data following FAIR principles. Within the first project phase from 2018-2021, a metadata scheme adapted to linguistic needs and processes for data preparation for archiving has been developed. SWISSUbase was launched in July 2022 as a national data repository with an adapted linguistic metadata schema that draws on the CLARIN CMDI and the META-SHARE schema. This schema will be interoperable with the Virtual Language Observatory.

3 The UpLORD project and its lines of action (2023-2024)

3.1 Rationale

The ORD practices targeted by the UpLORD project are those summarized by the FAIR principles for data management (cf. Wilkinson et al. 2016): *Findable, Accessible, Interoperable* and *Reusable*. At the same time, one main limit of Open research data is that not all data sets may fully adhere to the FAIR principles. This is the case of sensitive data, such as personal data which must be protected, or data having copyright and/or intellectual property issues. This type of data requires a special management.

In this context, we identified several gaps regarding the current situation in Switzerland which are now addressed the UpLORD project: (1) the provision and the usability of the individual corpus platforms across Switzerland: their drawback is that they do not satisfy the Interoperability principle, (2) the status of several linguistic corpora: most of them do not satisfy all four FAIR principles, (3) the lack of meaningful metadata and of infrastructures to manage a diversity of annotations of linguistic data, this being linked to the Findable and the Reusable principles, (4) the proper management of sensitive data, informed consent, copyright and intellectual property issues, which are necessary so that the sets of data adhere to the FAIR requirements, (5) the metadata schema on SWISSUbase which must be adapted and amended to satisfy the Findable principles, (6) uploading workflows on SWISSUbase, (7) the quality control and data curation in SWISSUbase to satisfy the Reusable principles, (8) the setting of ethical standards for best practices and frames of mind in linguistic data management and collaboration, as well as the lack of training for the target scientific communities to adopt these standards.

3.2 Lines of action

To fill in the above-mentioned gaps, the UpLORD project carries the following actions:

First, the implementation of the LCP@LiRI at the national level following the FAIR principles fills in gaps (1) and (2). As a first step, a group of three corpora of oral French are being integrated to the corpus platform which will be rebranded then as LCP@CLARIN-CH. In a second step, other corpora provided by the CLARIN-CH community will be integrated. There are numerous Swiss corpora (monolingual and multilingual) which lack of harmonized and standardized formats. The integration of these corpora is not only handled as a single case transformation problem, but general solutions for data conversions must be implemented. More precisely, we expect to receive common formats of corpora like XML-TEI (various flavours, e.g. DTABf by BBAW), CoLLN, XML style verticalized text formats, CSV, but also database outputs such as those of Swissdox@LiRI, etc.), thus converters have to be built depending on the actual needs. This will significatively improve discovery, access, integration, usability and reusability of corpora according to FAIR principles, as well as simplify re-formatting, assembling, harmonizing and standardizing the data and metadata.

Second, to fill in gap (3), a software application called VIAN-DH@LiRI is being developed for modelling complex annotation schemes that LCP@LiRI has to offer. Data processed within VIAN-DH is complex interactional data consisting of verbal, paraverbal and non-verbal annotation levels. Therefore, linear and token-based annotation models can't be used. This problem is similar to syntactic annotations, e.g. of phrase structures where hierarchical relations need to be modelled. So, by implementing VIAN-DH, these issues need to be solved and the data schema and the necessary query language will be developed. The data schema, realized as relational Postgres database, uses token and time based alignments. In addition, we will evaluate further test cases to develop a sustainable and flexible infrastructure with regards to annotation. The project will collect further complex annotation models via the CLARIN-CH partners as well as via other clients LiRI is working with. It is then evaluated whether the LCP@LiRI can also map these. This process goes hand in hand with the development of best practices to show researchers how to deal with complex annotations.

Third, to fill in gaps (4)-(6), the already implemented modular linguistic metadata schema of SWISSUbase will be specified and adapted to the needs that are not included for the moment (e.g. for neurolinguistic and experimental data). When it comes to publishing and archiving data, uploads on SWISSUbase are only possible through a web-based GUI-interface so far. To upgrade the usability of SWISSUbase, it should be also possible to provide a workflow via an API. In addition, easy workflows between LCP@CLARIN-CH and VIAN-DH will be implemented.

Fourth, to fill in gap (7) regarding quality control, through this project, we set up national working groups whose role is to develop specific metadata for various types of linguistic data (e.g., socio-linguistics data, psycholinguistics experimental, neurolinguistics data, conversational analysis data, lexicography data, computational data, acquisitional data, multimodal data) that satisfy the FAIR principles. To ensure the quality control of linguistic data, data curation is required, which is very time consuming. These services are provided at UZH, by both the Zurich University Library and LiRI. UZH provides data curation for all CLARIN-CH members and focuses on metadata quality, but a network of liaisons at the national level, across all disciplines using language-related data, still must be established to improve data quality and to disseminate information about standards and best practice in handling

metadata, as well as to overcome disciplinary boundaries. Informing and forming the scientific community upstream is crucial as it is impossible to improve data quality at the end of the data lifecycle. Researchers have to be aware that data quality issues start early in the project, already in the planning phase. For this, the national working groups planned in this project will focus on data quality issues with respect to awareness building for the whole data life cycle.

Fifth, to fill in gap (8) about setting ethical standards for best practices and providing training, this project will develop showcases, best practices and good habits for data management according for FAIR principles (such as, planning and writing data management plans, using sustainable file formats, setting solid file-naming conventions, finding data storage backup schemes, creating informative metadata and documentation) and will promote the ORD practice in the CLARIN-CH and other scientific communities. A special attention is given to finding solutions that overcome disciplinary boundaries.

4 Conclusion

The services proposed by the Swiss ecosystem of infrastructures provide access to and reuse of linguistic research data in Switzerland and abroad. Through the large array of partners represented by the CLARIN-CH consortium, the ecosystem of infrastructures benefits of numerous collaborations at different levels: (i) within the target national communities, (ii) between the two service providers and the target scientific communities, (iii) with the applicants of two other complementary ORD proposals. The UpLORD project focuses on upgrading workflows and interoperability of existing infrastructure services, establishing working groups on the national level, documenting and promoting best practices, raising awareness and training about ORD practices in the context of teaching, research and publishing, and building a robust practice of data curation. In sum, the Swiss FAIR-compliant ecosystem of infrastructures shares the vision of the European CLARIN infrastructure with respect to language resources, language technology and Open Science.

References

Mattern, E. (2022). The Linguistic Data Life Cycle, Sustainability of Data, and Principles of Solid Data Management. In *The Open Handbook of Linguistic Data Management*. (2022). Berez-Kroeker, A. L., McDonnell, B., Koller, E., Collister, L. B., (eds.), pp. 61-71. Cambridge MA: MIT Press. 10.7551/mitpress/12200.001.0001

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. and Bouwman, J. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18.

The ACoDe Project: Creating a Dementia Corpus for Icelandic

Elena Callegari
University of Iceland
Reykjavík, Iceland
ecallegari@hi.is

Agnes Sólmundsdóttir
University of Iceland
Reykjavík, Iceland
ags46@hi.is

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

Abstract

We are creating the very first Dementia corpus for the Icelandic language. Our corpus will contain manually transcribed speech samples elicited from individuals of Icelandic nationality who are aged 60 to 80 and who are suffering from various degrees of Alzheimer's disease. In this paper, we describe our speech elicitation protocol, how we collect the data and how we are transcribing the samples. By sharing our methodology, we hope to spark interest in cross-linguistic research collaborations to develop comparable corpora for languages other than Icelandic.

1 Alzheimer's Disease in Iceland

Alzheimer's disease (henceforth, *AD*) is a type of neurodegenerative disease that causes a progressive decline in cognitive faculties such as memory, decision making and language; around 25 million individuals across the world suffer from AD (Qiu et al., 2022). In Iceland, AD is a particularly pressing concern: according to a 2016 study (Jakobsdottir et al., 2016), people of Icelandic heritage are more likely than other European populations to carry a genetic mutation that results in a greater risk of developing Alzheimer's disease. Moreover, according to data released by the World Health Organization for the year 2019 [1], in Iceland, AD and other dementias were the top cause of death for women and the second cause of death for men. There is currently no cure for AD: there are only therapies that can treat its symptoms and possibly slow its progression. A timely Alzheimers diagnosis provides patients with a better chance of benefiting from existing treatments, with the possibility of accessing support systems and with more time to make plans for the future (Rasmussen and Langerman, 2019); it is, therefore, essential to diagnose this condition as soon as possible. The main procedures currently available to diagnose AD include cognitive tests in combination with PET or MRI, and/or the sampling of cerebrospinal fluid by means of lumbar punctures. These procedures are costly and have long waiting times. This results in delayed diagnoses but also in greater difficulties in monitoring the evolution of the pathology over time.

2 ACoDe: Developing Clinical Speech Analysis for Icelandic

AD affects what we say and how we say it. Multiple studies have shown that individuals suffering from AD exhibit difficulties with word retrieval (Croisile et al., 1996, Kavé and Dassa, 2018), produce fewer information units and content words (Ahmed et al., 2013, Croisile et al., 1996, Kavé and Dassa, 2018), and use more pronouns than healthy age-matched controls (Kavé and Dassa, 2018). Changes to language are already detectable when individuals are diagnosed with Mild Cognitive Impairment (Kavé and Dassa, 2018), a stage of the disease that can occur up to 8 years before the onset of mild Alzheimer's dementia. Spoken language can thus offer a universal and accessible means for measuring neurological health and diagnosing early-stage AD. Indeed, automatic feature extraction and analysis of spoken language for clinical purposes have been attempted before, with remarkable results (Fraser et al., 2014, Peintner et al., 2008, i.a.). Nothing of the sort however currently exists for Icelandic. The ACoDe project ("**A**ssessing **Co**gnitive **De**cline using automatic language analysis") seeks to remedy this gap: our goal

[1] https://data.who.int/countries/352

is to develop software specifically designed for the Icelandic language, aimed at diagnosing Alzheimers disease through automated language analysis. In order to do that, we are collecting speech samples from Icelandic individuals suffering from various stages of AD as well as from healthy, age-matched individuals. Once all participants have been tested, we will train different classifiers on the resulting dataset, to determine whether the classifiers can distinguish between patients and controls, between different stages of AD, and with which accuracy. Our plan is to release the transcriptions in the form of a publicly accessible dataset, so that any researcher working on AD, clinical applications for NLP, or both, may also make use of the data we are collecting. In this paper we describe the speech sample collection process, how we are transcribing the data and how we plan on anonymizing it.

Although our corpus is still in the development phase, we are excited to share our methodology as it is our hope that this will act as a catalyst for future cross-linguistic collaborative research efforts. We are actively seeking partnerships for a pan-European initiative to develop comparable corpora for languages other than Icelandic. We believe that a pan-European project of this type would provide a breakthrough understanding of the effects of Alzheimer's disease on language. For example, analyzing the speech and language patterns of individuals with AD in multiple languages would help confirm the universality or specificity of certain diagnostic markers, possibly leading to breakthroughs in both diagnosis and treatment. Moreover, by working collaboratively on a more global scale, resources (both human and computational) could be shared more efficiently, speeding up the time it takes to compile sufficient data and thus to reach meaningful conclusions.

2.1 Innovative Value

Ours will be the very first dementia dataset for Icelandic. Most of the existing work on clinical feature extraction from language has been done for English (Williams et al., 2021). English also dominates clinical language datasets (MacWhinney et al., 2011), so understanding the extent to which language deterioration due to brain disease generalizes across languages is of great interest. In this respect, Icelandic is an excellent addition to the research effort: Icelandic has a complex inflectional morphology, V2 in embedded clauses (only Yiddish also has embedded V2), as well as other linguistically interesting properties (Thráinsson, 2007). Thus studying how AD affects a language like Icelandic provides invaluable information to better understand how AD affects language more in general.

3 Status & Participants

The ACoDe project officially started in mid 2022. Up until now, we have collected speech samples from more than 50 individuals; we expect to complete the speech collection process by the end of 2024.

We plan on recruiting a total of 120 individuals: 30 patients suffering from Mild Alzheimer's Dementia (MD), 30 patients suffering from Mild Cognitive Decline (MCD) and 30 patients suffering from Subjective Cognitive Decline (SCD). Finally, we will also include 30 healthy controls. There will thus be 3 diagnostic groups and 1 control group, for a total of 4 research groups. Patients with SCD complain of memory problems and overall reduced cognitive abilities, but the extent of these disorders is not such as to be detected by standardized cognitive tests (hence the label *subjective*). Several studies have shown an association between SCD and an increased risk of developing various forms of dementia (Jonker et al., 2000, Geerlings et al., 1999, Jessen et al., 2014, Jessen et al., 2010), hence our interest in this condition. The diagnosis of MCI, MD or SCD is made by qualified clinicians from the Memory Clinic in Reykjavík. All participants will be between the age of 60 and 80. The exclusion criteria (for both patients and controls) are a primary diagnosis of depression of moderate or severe degree, bipolar disorder, schizophrenia, a previous physical brain injury, a neurological disorder or other serious medical condition, a personal history of drug addiction within the past 20 years, issues with alcohol addiction within the past 20 years, the use of antidepressants and the use of benzodiazepine-based sleep medications. To avoid potential confounding factors due to the knowledge of a second language, we are also only accepting individuals who are monolingual speakers of Icelandic. Our study received approval from the Icelandic Research Ethics Committee (*Vísindasiðanefnd*) in September 2021.

We are committed to achieving gender balance across study groups whenever this is possible. In the

control group, we have successfully attained an equal gender distribution, with 15 males and 15 females participating in our study. Achieving such balance is more challenging in the patient groups, as the pool of possible participants is much smaller.

3.1 Speech Elicitation Protocol

Each participant is asked to describe in detail: (i) the "picnic scene" by The Arizona Alzheimer's Disease Center. This is a black-and-white depiction of a picnic by the lake; (ii) how they would plan a trip to Akureyri, a city in the north of Iceland; (iii) their childhood home. We decided to include more than the traditional picture-description task, used in many studies on AD, because of evidence that picture-description tasks may not accurately reflect the conversational abilities of individuals with AD (Sajjadi et al., 2012). We chose to include the planning-a-trip kind of narrative task following (Harris et al., 2008), who included a "Plan a trip to New York" question in their own study. According to (Harris et al., 2008), the planning-of-a-trip scenario is complex enough to detect differences between healthy controls and individuals with cognitive decline. This kind of narrative is also effective at engaging episodic memory, which is impaired in individuals suffering from AD (Economou et al., 2016). Finally, we chose to ask participants to describe their childhood home because we reasoned that this question is likely to elicit long responses, minimizing the need for the interviewer to prompt the interviewee with additional follow-up questions to increase the total length of the speech sample. Note also that the description of the childhood home focuses on past events, the trip-planning pertains to future activities, and the description of the picnic scene is not anchored to a specific time. Therefore, this approach could also offer insights into how the disease affects cognitive processing of temporal events and the linguistic expression of time.

The order in which the three main prompts are presented is rotated across participants to mitigate the effect of fatigue on verbal performance. During interviews, participants are encouraged to speak freely and uninterrupted while being audio-recorded. The interviewer uses nonverbal cues and encouraging feedback to make the conversation feel as natural as possible. The goal is to elicit 15 minutes of spoken recording from each participant, and if necessary, pre-decided follow-up questions are asked to elicit longer speech or keep the discussion on topic. The interviewer generally waits 10 seconds after the interviewee has finished speaking before asking follow-up questions. We strive to always ask the follow-up questions in the same order, so as to ensure that speech samples from different participants are maximally comparable. However, we may adjust the question order to maintain a natural conversation, especially if an answer to any of the sub-questions has already been provided earlier.

3.2 Data Anonymization

Participants sign an informed consent at the beginning of the interview which states that all of the participants' identifiable personal information is confidential. Each participant is assigned a research number under which all their data produced in the study is stored. The only link between participant name and research number is kept in an encrypted file which can only be accessed by the interviewer and the researchers at the Memory Clinic who conduct the EEG and cognitive tests. Despite all participants' data being stored anonymously, some identifiable and traceable information can appear during interviews. This is to be expected particularly during the prompt on the participants' childhood home, which often leads to descriptions of family members, and to mentions to schools, specific places and organizations. As Iceland is a fairly low-populated community, this information can in principle reveal the identity of the participant. All information that is deemed to be traceable will therefore be redacted from the transcription dataset before publication, in a way that makes the interviewee unidentifiable while still providing equivalent lexical information needed for language analysis. There are several ways this can be done. One method, recommended by (Aldridge et al., 2010), is to replace potential identifiers such as names, places or organizations with unique identifiers (e.g., pseudonyms), rather than anonymous placeholders (e.g., Person Name). We intend to follow the anonymization guidelines detailed in (Francopoulo and Schaub, 2020). However, owing to Iceland's small population, we must also modify certain phrases that are not specified in the guidelines but could make individuals identifiable in smaller communities. Such elements may include, but are not limited to, names of schools, cities, towns, regions, individuals,

and specific dates. We intend to follow the method by which original items are replaced with pseudonyms -or made-up dates in the case of dates-, therefore retaining all grammatical information while protecting anonymity. For example, the fragment sentence in 1, which discusses a local Icelandic school, would be published as 2, where the original school name has been replaced with a pseudonym, keeping the grammatical properties of the original sentence.

(1) ***Já ég var í sa- sama skólanum ee í Langholtsskóla***
`Yes I was in sa- same school uh in Langholtsskóli-DAT.`
'Yes I went to the same school, Langholtsskóli (A school in Reykjavík).'

(2) ***Já ég var í sa- sama skólanum ee í Borgarskóla***
`Yes I was in sa- same school uh in Borgarskóli-DAT.`
'Yes I went to the same school, The City School (A fake school that doesn't exist).'

4 Transcription Protocol

Using transcription methods and guidelines from the *Linguistic Data Consortium at the University of Pennsylvania* (hence, LDC), we generate manual text transcriptions from the recorded speech samples (Glenn et al., 2010). The transcriptions are made using a standard text processor, such as Microsoft Word or TextPad, and exported to plain text files (.txt). The transcriptions contain speech from both speakers, i.e. interviewer and interviewee, and accurately annotate any interjections or overlaps, providing detailed transcriptions of the conversations as a whole. The transcriptions are verbatim and orthographic using standard Icelandic spelling. Filled pauses, false starts, repeated words, repairs, restarts, partial words, spoonerisms, speech errors and speaker noises are all marked and annotated in accordance with the transcription protocol. We follow the LDC guidelines as much as possible with some modifications for Icelandic. These adjustments primarily involve Icelandic discourse particles, which differ from those in English. For instance, we created a list of Icelandic-specific particles, including "uu", "ömm", "sko", and "hérna". The word "hérna" is noteworthy, as it serves as both a lexical word, an adverb of place meaning "here", and a planning marker used to indicate hesitation or to maintain a speaker's turn in a conversation, similar to the English particle "uhm" (Hilmisdóttir, 2011). We therefore annotate this word to differentiate between the two meanings as shown in 3 and 4.

(3) Adverb of Place

Hérna situr par á teppi.
`Here sits couple on blanket.`
'Here is a couple sitting on a blanket.'

(4) Planning Marker

Og *hérna* það var bara *hérna* mjög gaman.
`And *uhm* that was just *uhm* very fun.`
'And that was just very fun.'

5 Format

All transcriptions of the speech samples resulting from our project will be made publicly available in the form of a CC-BY 4.0-license corpus, which will be distributed in TEI-conformant format and will be accessible to everyone on the Icelandic CLARIN repository. The corpus will only include the transcriptions, not the audio files, as it is much harder to preserve anonymity with audio files. The released version will include annotation that builds on several other Icelandic CLARIN resources, released via projects that have been carried out in recent years, most notably the Icelandic Language Technology Programme Nikulásdóttir et al., 2020, and build on the experience and protocols accumulated within these projects, ensuring interoperability between systems and readiness of the human resources available in Iceland for future work. Our open-source policy and standardized packaging of our data will encourage the use of our output in future R&D projects across academia and industry.

References

Ahmed, S., Haigh, A.-M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven alzheimers disease. *Brain*, *136*(12), 3727–3737.

Aldridge, J., Medina, J., & Ralphs, R. (2010). The problem of proliferation: Guidelines for improving the security of qualitative data in a digital age. *Research Ethics*, *6*(1), 3–9.

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and language*, *53*(1), 1–19.

Economou, A., Routsis, C., & Papageorgiou, S. G. (2016). Episodic memory in alzheimer disease, frontotemporal dementia, and dementia with lewy bodies/parkinson disease dementia. *Alzheimer Disease & Associated Disorders*, *30*(1), 47–52.

Francopoulo, G., & Schaub, L.-P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. *workshop on Legal and Ethical Issues (Legal2020)*, 9–14. https://hal.science/hal-02939437

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *cortex*, *55*, 43–60.

Geerlings, M. I., Jonker, C., Bouter, L. M., Adèr, H. J., & Schmand, B. (1999). Association between memory complaints and incident alzheimers disease in elderly people with normal baseline cognition. *American Journal of Psychiatry*, *156*(4), 531–537.

Glenn, M. L., Strassel, S. M., Lee, H., Maeda, K., Zakhary, R., & Li, X. (2010). Transcription methods for consistency, volume and efficiency. *LREC*.

Harris, J. L., Kiran, S., Marquardt, T. P., & Fleming, V. B. (2008). Communication wellness check-upl': Age-related changes in communicative abilities. *Aphasiology*, *22*(7-8), 813–825.

Hilmisdóttir, H. (2011). Giving a tone of determination: The interactional functions of nú as a tone particle in icelandic conversation. *Journal of Pragmatics*, *43*(1), 261–287.

Jakobsdottir, J., van der Lee, S. J., Bis, J. C., Chouraki, V., Li-Kroeger, D., Yamamoto, S., Grove, M. L., Naj, A., Vronskaya, M., Salazar, J. L., et al. (2016). Rare functional variant in tm2d3 is associated with late-onset alzheimer's disease. *PLoS genetics*, *12*(10), e1006327.

Jessen, F., Amariglio, R. E., Van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., Dubois, B., Dufouil, C., Ellis, K. A., Van Der Flier, W. M., et al. (2014). A conceptual framework for research on subjective cognitive decline in preclinical alzheimer's disease. *Alzheimer's & dementia*, *10*(6), 844–852.

Jessen, F., Wiese, B., Bachmann, C., Eifflaender-Gorfer, S., Haller, F., Kölsch, H., Luck, T., Mösch, E., van den Bussche, H., Wagner, M., et al. (2010). Prediction of dementia by subjective memory impairment: Effects of severity and temporal association with cognitive impairment. *Archives of general psychiatry*, *67*(4), 414–422.

Jonker, C., Geerlings, M. I., & Schmand, B. (2000). Are memory complaints predictive for dementia? a review of clinical and population-based studies. *International journal of geriatric psychiatry*, *15*(11), 983–991.

Kavé, G., & Dassa, A. (2018). Severity of alzheimers disease and language features in picture descriptions. *Aphasiology*, *32*(1), 27–40.

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). Aphasiabank: Methods for studying discourse. *Aphasiology*, *25*(11), 1286–1307.

Nikulásdóttir, A., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Steingrmsson, S. (2020). Language technology programme for icelandic 2019-2023. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3414–3422.

Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M. L. G., & Ogar, J. (2008). Learning diagnostic models using speech and language measures. *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 4648–4651.

Qiu, C., Kivipelto, M., & Von Strauss, E. (2022). Epidemiology of alzheimer's disease: Occurrence, determinants, and strategies toward intervention. *Dialogues in clinical neuroscience*.

Rasmussen, J., & Langerman, H. (2019). Alzheimers disease–why we need early diagnosis. *Degenerative neurological and neuromuscular disease*, 123–130.

Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, *26*(6), 847–866.

Thráinsson, H. (2007). *The syntax of Icelandic*. Cambridge University Press.

Williams, E., McAuliffe, M., & Theys, C. (2021). Language changes in alzheimers disease: A systematic review of verb processing. *Brain and Language*, *223*, 105041.

A Continous Integration (CI) Workflow for Quality Assurance Checks for Corpora of Multimodal Interaction

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

André Frank Krause
University of Duisburg-Essen, Germany
andre.krause@uni-due.de

Karola Pitsch
University of Duisburg-Essen, Germany
karola.pitsch@uni-due.de

Abstract

When creating multimodal language resources, ensuring a high data quality standard is central. While this is true for every language resource, multimodal resources pose additional challenges for quality assurance, like audio-visual material synced with annotations and transcriptions and sensor data aligned with speech or other modes of communication such as gestures. Making use of existing data quality assurance frameworks and GitLab CI we present best practices for the sustainable quality checks of multimodal and multisensorial corpora established in the MuMo-Corp project situated in the context of Interactional Linguistics and Conversation Analysis.

1 Introduction

When creating corpora of language use and social interaction which can be made available for other researchers through institutional repositories, ensuring a high data quality standard is central. When preparing a corpus, researchers can draw on some well established practices in the CLARIN ecosystem from Corpus Linguistics, Spoken Language Corpus Linguistics/Pragmatics and Language Documentation (e. g. Schmidt, 2016, Rühlemann, 2018, Arkhangelskiy et al., 2020, Ferger and Jettka, 2021)[1]. Yet, we encounter a range of shortcomings of these approaches once we attempt to prepare a corpus of multimodal human-robot-interaction (e.g. Pitsch, 2020) which has grown successively over a range of years.

In this paper, we focus on one specific aspect of corpus management, the quality checking of multimodal and multisensorial transcriptions and annotations as they emerge from timeline- and XML-based tools such as ELAN (Sloetjes, 2014) which we have developed within the data-reuse project "MuMo-Corp"[2]. When transferring the existing data into a structured and reusable corpus (Hedeland, 2020), we encountered the following requirements which the quality checking should meet: A) Given the multimodal nature of the data, the quality checks need on the one hand to respect standardized conventions (e.g. GAT 2) and on the other hand to be adaptable to project-specific annotation conventions once bodily phenomena are involved. B) To convert the corpus data to different formats in order to integrate them in existing corpus tools providing GUI access (e.g. AGD, ZuMult), to use them for automated analysis (e.g. using R) and for storage in an institutional longterm repository (e.g. the IDS Repository which is a member of CLARIN), the quality checks and fixes need to be re-introduced in the original transcription/ annotation files (such as ELAN). C) We wanted the quality checks to be, to a large extent, automated

[1] Previous efforts by the data centres AGD (the Archive for Spoken German) and the HZSK (the Hamburg Centre for Language Corpora), which are both CLARIN B Centres, as well as the projects INEL and QUEST cooperating with the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation are crucial to these developments.

[2] https://www.uni-due.de/kowi/mukom/mumocorp

while, at the same time, maintaining human readability for manual control. D) The workflows should be reproducible and reusable with other corpora and also during an ongoing project.

When reviewing existing practices for quality checking of corpora, we encountered some shortcomings, such as limited possibilities for automation and reproducibility while maintaining human readability, or the practice of performing quality checks and fixes on dataframes which have been exported from the original transcriptions and which cannot be re-introduced in the original transcriptions. Therefore, in this paper we explore novel workflows which attempt to answer the following questions: How can practical measures for quality assurance on multimodal corpora be carried out in a sustainable way? How can they be easily adapted by further projects in the future?

Using the example of the MuMoCorp project[3] focusing on multimodal human-robot-interaction and the design and use of novel technologies we will introduce best practices, reusable workflows and tools to be adapted by further projects faced with the same challenges.

2 Transferring MuMoCorp Data into a Structured and Coherent Corpus for Analysis, Reuse and Archiving

Within the data reuse project "MuMoCorp", we were faced with a collection of audiovisual recordings, robot logfiles and motion capture data with accompanying plots and animations, timeline-based transcriptions (XML, using ELAN) and multimodal annotations of several consecutive studies of human-robot-interaction in the scenario of a museum guide robot. This extremely rich collection has emerged over a period of 10 years within an interdisciplinary team of researchers and with funding from different projects with their respective focused research questions (e.g. Pitsch, 2020). These files were stored in a systematic file structure which has grown over the years with a range of successive expansions, naturally leading to slight inconsistencies in e.g. file naming. Our goal for the data reuse project consists in transferring this data into a documented, structured and coherent multimodal interactional corpus that is machine-readable (cf. Hedeland, 2020) and which can be used for analysis (both automatic and manual), for reuse in different research contexts and for longterm archiving. When preparing the corpus data, different tasks at different stages of data processing become relevant. Based on the proposed stages of Corpus Initialisation, Data Curation and Corpus Integration in Hedeland and Ferger, 2020 we created an extended overview of these tasks (Fig. 1) and asked ourselves which competences (subject specific / technical) are required to solve them (represented in the color-coding).

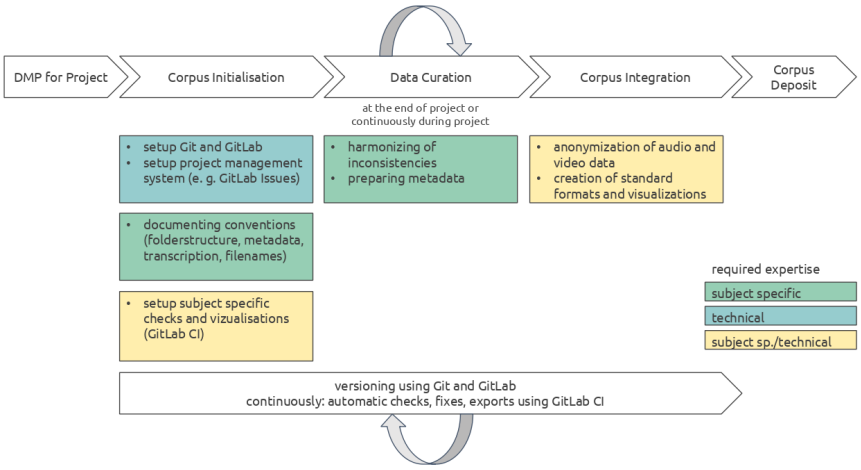


Figure 1: Stages and exemplary tasks for multimodal corpus projects

When preparing the corpus data, we not only carried out each task separately, but aimed at developing workflows that can be reused with further projects at the intersection of Conversation Analysis and

[3]https://www.uni-due.de/kowi/mukom/mumocorp

Corpus Linguistics / Corpus Pragmatics. With this perspective, to check the existing transcriptions and annotations and to harmonise them, what we call Data Curation here, is not only relevant at the end of a research project when curating an already existing collection of data and annotations. Rather, the same tasks can also be carried out continuously during ongoing research projects. As shown in Hedeland and Ferger, 2020 by performing quality assurance measures continuously during a project rather than once before publication decreases the time and effort necessary for these measures.

3 Best practices for Quality Assurance of Multimodal Corpora

In what follows, we present new workflows for creating coherent and sustainable multimodal interactional corpora as we have developed them within the data reuse project described in the previous sections. A live example including all scripts referred to here that can be adapted to other projects can be found in the Git-Repository "Elan CI Checks Example"[4] in the public GitLab group "MuMoCorp Open Access" which collects the resources established within the data reuse project.

3.1 Continuous quality control in GitLab using GitLab CI

The requirements to apply normalizations and fixes to the original transcription sources and to reuse and/or apply quality checks continuously to the data are, on a technical level, best met with methods for continuous quality control. Therefore, we propose a workflow that builds on and goes beyond existing methods in continuous quality control (as in Hedeland, 2020, Ferger and Jettka, 2021) by using GitLab with its Continuous Integration (CI) functionality. Other examples of CI functionality are GitHub Actions or Bitbucket Pipelines, but though the setup can be easily adapted to these platforms, a university-hosted GitLab instance allows the data to be stored on university infrastructure and not proprietary servers, which is especially important when dealing with research data containing (even anonymized) personal information. Such CI functionalities have originally been created for automated deployment of software but also offer great advantages for data quality checks (see e.g. Herrmann et al., 2021, Erjavec and Kopp, 2022 Cyra et al., 2022). It automatically and continuously applies specified scripts on each change of the data. GitLab CI functionality also facilitates the use of different technologies in one workflow. To leverage it, the research data needs to be versioned using git and GitLab (see e.g. Cyra et al., 2022). The components for quality checking which we applied and/or developed within the data reuse project (see section 3.2 and 3.3) have been integrated in this GitLab CI workflow.

3.2 Corpus quality checks for interactional ELAN files and fixing inconsistencies automatically

The transcriptions with which we were faced in the data reuse project have been created in ELAN (Sloetjes, 2014), so that our work of quality checking could build on functionalities offered by the Corpus Services framework (Ferger et al., 2020, Hedeland and Ferger, 2020). As this was initially developed to work mainly with EXMARaLDA files, some further developments for ELAN files had been added in other contexts (see e.g. Arkhangelskiy et al., 2020). In the data reuse project, we make use (with some small enhancements) of the already existing *ELANValidatorChecker* (checks if the ELAN XML file is valid), the *ELANFileReferenceChecker* (checks if the linked audio and video files exist), *ELANTranscriptionChecker* (checks transcriptions according to standardized conventions, in our case GAT, as stated in our requirements in section 1), *ELANAnnotationChecker* (checks if annotation tiers adhere to annotation conventions, which is especially important for our coded interactional annotations). At the same time, we added new functionalities to the framework, such as an automated conversion of ELAN files to EXMARaLDA files.

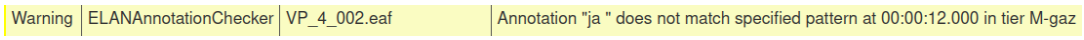


Figure 2: Example of error list display from the Corpus Services Framework in GitLab CI

An advantage of Corpus Services is the generation of a human-readable and sortable output file which

[4] https://git.uni-due.de/mumocorp-open-access/elan-git-example

lists different inconsistencies that need to be repaired manually. When this output file is treated as an 'artifact' in GitLab CI, it is displayed and formatted conveniently (figure 2)[5].

A further benefit of using GitLab CI is the possibility to integrate different programming languages in the workflow, so that already existing scripts (e.g. in R or Python) could easily be added and new scripts could be created in a convenient way. This includes scripts for checking transcriptions and annotations, such as identifying superfluous whitespaces in the wrong locations or the name of annotation tiers using R with regular expressions. The existing R scripts[6] were adapted to automatically fix the identified inconsistencies. This could be done since the ELAN XML were used as sources directly and could be written back in the ELAN XML format after application of the fixes.

3.3 Exporting further formats from interactional ELAN files

As the corpus data should be usable to be integrated in existing corpus tools providing GUI access (e.g. AGD, ZuMult), to be used for automated analysis (e.g. using R) and for storage in an institutional longterm repository (e.g. the IDS Repository) the quality checks and ensuing fixes have been re-introduced in the original transcription/annotation files (here: ELAN) leveraging the GitLab CI setup. Thus keeping measures for quality assurance in the source files, different export functionalities for building tables in CSV format, for building R dataframes or TEI have been created on the basis of the enhanced ELAN XML source files and added to the GitLab CI setup as well. Thus, with each change of the original file, these exports are regenerated and can easily be downloaded.

4 Conclusion

Sustainable quality assurance of research data is crucial for complex and unique data collections, such as multimodal and multisensorial corpora of interaction. While there is a lot of research and work the interactional linguistics and conversation analysis can build upon, the specific data type required some adaptions, such as checking annotations depending on their communication mode as in example 2. By sharing our setup and scripts we hope to enable further developments and collaboration on these tasks.

Acknowledgments

This project was financed by the Volkswagenstiftung (grant number 90886, PI: Karola Pitsch).

References

Arkhangelskiy, T., Hedeland, H., & Riaposov, A. (2020). Evaluating and assuring research data quality for audiovisual annotated language data. In C. Navarretta & M. Eskevich. (Eds.), *Clarin Annual Conference Proceedings, 2020. ISSN 2773-2177 (online)* (pp. 131–135). Virtual Edition, 2020.

Cyra, M. A., Politze, M., & Timm, H. (2022). A push for better RDM: Erfahrungsbericht aus dem Einsatz von git für Forschungsdaten. *Bausteine Forschungsdatenmanagement*, (2).

Erjavec, T., & Kopp, M. (2022). TEI and Git in ParlaMint: Collaborative Development of Language Resources. In T. Erjavec & M. Eskevich (Eds.), *Clarin Annual Conference Proceedings, 2022. ISSN 2773-2177 (online)* (pp. 57–60). Prague, Czechia, 2022.

Ferger, A., Hedeland, H., Jettka, D., & Pirinen, T. (2020). Corpus Services (version 1.0). https://doi.org/10.5281/zenodo.4725655

Ferger, A., & Jettka, D. (2021). Seamless integration of continuous quality control and research data management for indigenous language resources. In M. Monachini & M. Eskevich (Eds.), *Clarin Annual Conference Proceedings, 2021. ISSN 2773-2177 (online)* (p. 95). Virtual Edition, 2021.

Hedeland, H. (2020). Towards comprehensive definitions of data quality for audiovisual annotated language resources. In C. Navarretta & M. Eskevich. (Eds.), *Clarin Annual Conference Proceedings, 2020. ISSN 2773-2177 (online)* (pp. 93–103). Virtual Edition, 2020.

[5]The realtime error list of the example project can be viewed here http://git.uni-due.de/mumocorp-open-access/elan-git-example/-/jobs/artifacts/main/file/report-output.html?job=corpus-services_check

[6]Especially some Regular Expressions were developed in a workshop with Rühlemann, cf. Rühlemann, 2020

Hedeland, H., & Ferger, A. (2020). Towards Continuous Quality Control for Spoken Language Corpora. *International Journal of Digital Curation*, *15*(1). https://doi.org/https://doi.org/10.2218/ijdc.v15i1.601

Herrmann, F., Pietsch, C., & Cimiano, P. (2021). Conquaire Infrastructure for Continuous Quality Control. In C. Pietsch, P. Cimiano, C. Wiljes, & B. University (Eds.), *Studies in analytical reproducibility: The conquaire project* (pp. 17–27).

Pitsch, K. (2020). Answering a robot's questions: Participation dynamics of adult-child-groups in encounters with a museum guide robot [Postprint]. https://doi.org/10.3917/res.220.0113

Rühlemann, C. (2018). *Corpus Linguistics for Pragmatics: A guide for research* (1st ed.). Routledge. https://doi.org/10.4324/9780429451072

Rühlemann, C. (2020). *Visual Linguistics with R: A practical introduction to quantitative Interactional Linguistics*. John Benjamins Publishing Company. https://doi.org/10.1075/z.228

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*, *31*(1), 105–132.

Sloetjes, H. (2014). ELAN: Multimedia annotation application. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *Handbook on Corpus Phonology* (pp. 305–320). Oxford University Press.

Swissdox@LiRI – a large database of media articles made accessible to researchers

Johannes Graën, Igor Mustač, Nikolina Rajović,
Jonathan Schaber, Gerold Schneider, Noah Bubenhofer
Linguistic Research Infrastructure
University of Zurich, Switzerland
first_name.last_name@linguistik.uzh.ch

Abstract

This article presents our efforts to make a large collection of Swiss newspaper articles available for research purposes. We describe the resource, detail the concept of financing and explain the application we built for researchers to obtain datasets from Swissdox@LiRI. To date, more than 250 users have compiled more than 1300 datasets with an average size of approximately 200 000 articles.

1 Introduction

The 'Schweizer Mediendatenbank AG' (SMD)[1] is a nonprofit joint venture of three big Swiss media groups with the purpose of collecting print and online publications, as well as TV subtitles, primarily by Swiss news agencies. They operate a database with the same name, SMD, which can be accessed free of charge by their partners and, by members of journalist organizations for a fee. Swissdox AG[2] is a subsidiary of SMD which offers fee-based access to the SMD database for everyone, in contrast to the limit access SMD provides. Monthly fees range from 110 CHF to 1200 CHF (before tax) and allow access to a limited number of publications (250 to 12 000 documents per month).[3] Special conditions are available for schools, libraries and universities. They can apply for a license with unlimited access. The web application 'Swissdox essentials' is the tool used in these cases to query the database and explore individual results. The application only shows the results that are ranked highest (with different ranking options), but does neither allow to browse all results nor to download more than a single document at once.

The main objective of uniting articles from different sources and compiling the SMD database is to create a comprehensive archive of press material for journalists with a view to improve quality in journalism.[4]

2 Related Work

The DeReKo platform (Deutsches Referenzkorpus) is a corpus platform for the German language, providing access to a vast collection of written texts, comprising also news articles besides other genres (Kupietz, 2010). It is still under active development at the Institut für Deutsche Sprache (IDS) in Mannheim, growing by approximately 200 millions of tokens per year. The texts are stored as XML and made available through custom applications like COSMAS II (Bodmer Mory, 2005) and KorAP (Bański et al., 2013). However, DeReKo is not update on a daily basis and also does not strive to offer comprehensive coverage of German news paper articles. Other corpora that include (German) news articles are e.g. the DWDS-corpus (Geyken, 2007) or the TIGER Treebank (Brants et al., 2004); but the aim of both of these corpora, similarly to DeReKo, is not current, timely-updated news coverage.

[1] https://smd.ch/

[2] https://swissdox.ch/

[3] https://swissdox.ch/abonnements/ (September 11, 2023)

[4] https://smd.ch/de/about (September 11, 2023)

3 Resource

Data collected in the SMD database consists of metadata and the actual articles which are encoded in a proprietary XML format. Metadata include the source of each article, its publication time, title or titles, and, optionally a rubric and regional mapping. Furthermore, information on its language, character count, technical origin (digitally provided vs. obtained from scanned images via OCR methods), and a document type of the respective source[5] are provided.

By running an analysis on a subset of the data, we found that several articles had been published more than once. In fact, associated media outlets have access to a common pool of articles, which are often published in all related media simultaneously.

Depending on the individual application, it might be required to know about duplicate articles. In other cases, e.g. linguistic data analysis, leaving duplicates in the data can lead to a bias.

We opted for a data model that pays respect to this observation and tries to countervail data duplication by normalizing the – otherwise flat – data into three categories:

1. The article **content**, which only comprises the article in XML format and an identifier derived from that. Since this identifier is a hash computed on the XML content, identical article contents will receive the same identifier.

2. The **article** itself comprising its content, the titles (dateline, heading and subheading), and the document type, character count and technical origin as metadata. For articles, we also derive an identifier based on the entirety of attributes such that two identical articles receive the same.

3. The **publication**, which includes the article with its content, the respective source medium, and the publication time. For online publications, the URL to the original publication is provided, as well as rubric and regional mapping when available. Analog to articles and their content, an identifier is derived from the data provided. Identical publications are suppressed.

We import between 5 000 and 6 000 new articles — i.e. non-normalized "flat" data — that we receive from SMD on a daily basis. Before actually importing the data into our relational database, we apply the above described normalization. We further disregard data points that contain flawed information, such as missing article contents (only headlines available), article without titles etc.

While Swissdox@LiRI mainly serves the purpose of providing recent news paper data (daily updates), we also received data in retrospective from 1911 on. However, the data volume becomes only substantial from the mid 1990s on, since only from then on more than 100 000 non-duplicated articles, i.e. what we call "publications", per year are available. Figure 3 shows the number of publications available per year in Swissdox@LiRI. In total, there are 25.6 millions publications available in Swissdox@LiRI; table 1 gives an overview of the shares of each language present in the data.

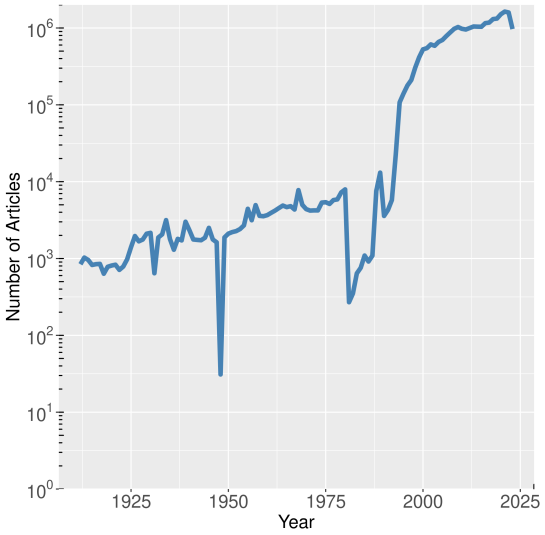


Figure 1: Publications/year in Swissdox@LiRI (note the log-scale).

[5] By means of a classifier defined by the "Press Database and Licensing Network" (https://www.pdln.info/ (September 11, 2023).

language	percentage
German	83.60
French	15.93
Italian	0.24
Rumantsch	0.16
English	0.07

Table 1: Shares of languages present in Swissdox@LiRI.

When users compile a dataset from Swissdox@LiRI (see Section 5), we produce a tabular format again and thus denormalize data into a similar state in which we retrieved it. Identical articles and article contents (e.g. with different headlines) can easily be identified using the above-mentioned identifiers.

4 Business Model

The main source of financing for the development of Swissdox@LiRI comes from supporting institutions (currently six universities and university libraries). In exchange, members of those institutions are granted access to the media database. Our Terms of Use require users to delete data retrieved from Swissdox@LiRI after a grace period of six months after closure of the corresponding research project. The actual articles cannot be disseminated, but derivatives thereof such as statistics, language models or example sentences. Data obtained from Swissdox@LiRI cannot be used commercially; a non-commercial license needs to be applied to any derived work.

In addition to supporting institutions, individual institutes and departments, research groups or other academic units can obtain individual licenses with different limits in terms of registered projects, accounts or number of articles that can be retrieved. When the total of license fees exceeds 180 000 CHF, individual supporter fees are reduced proportionately, as this sum is sufficient for us to maintenance of database and application.

5 Our Application

There are two ways to access the Swissdox@LiRI data: First, registered users can query the corpus by means of a web application where queries are entered in a form, or, secondly, they can access it via an API. We describe the web interface in the following. We use a simple variant of regular expressions as query language – however, we intend to incorporate the data from Swissdox@LiRI into another corpus tool, where a advanced query language will be available (see section 6. Users may also select the language (German, French, Italian, English), the time frame (we cover the past 26 years), the source (we have about 200 newspapers) and the media type (daily, weekly, online, etc.). The retrieval of matching articles from the whole corpus typically takes several minutes, therefore we implement a queuing system where users are alerted via email containing a download link with the compressed result set when the data is ready. They can then download it, for further processing on their own computer. Figure 2 shows the query interface, accessed in a web browser.

6 Use Cases

Many projects have been registered at Swissdox@LiRI since its launch in 2022; we list here a selection of publications that worked with data from this application. In their study "Conceptualizing Landscapes Through Language" Purves et al., 2023 examine how different languages and competence levels influence how people conceptualize landscapes. Ort et al., 2023 unraveled the development and spreading of "key subtopics and their evolution throughout the pandemic, and to identify key actors and their relationship with different aspects of the discourse around the pandemic." One project (Vamvas et al., 2023) fine-tuned a BERT language model (Devlin et al., 2018) on Swiss Media Text, while another "develops

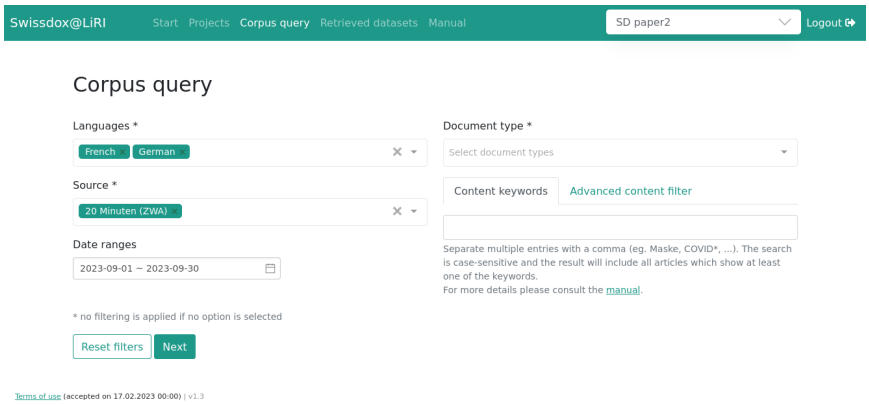


Figure 2: Web-based Query Interface to Swissdox@LiRI

an index of Climate Policy Risk using text-analysis techniques on a large number of Swiss media articles for the period 2000-2022" (Berthold, 2023).

7 Future Plans

Our future plans include the integration of automated text analysis, such as readability scores, compute word embeddings, create topic models on the fly, and offer visualizations. We will integrate Swissdox@LiRI into our corpus platform (Schaber et al., submitted) to offer advanced querying and analyze differences by metadata, including significance testing and time series analysis.

Queries for semantic analysis using word embeddings will allow the retrieval of related articles, automated summarization with BERT models, and passage retrieval with QA methods.

We also envisage to offer an interactive R and Python environment with ShinyR or Binder, allowing users to adapt and extend sample code and customize visualization, for the benefit of media science, political science, linguistics and historians alike.

In the current application, we will provide stable links for queries to facility the reproduction of analyses by other users, in line with the FAIR requirement of data being findable (Wilkinson et al., 2016). Individual articles can be changed or retracted after publication, which might impede the reproducibility of individual studies. The number of results obtained together with all query parameters, however, will be sufficient information in most cases to replicate a work based on Swissdox@LiRI.

References

Bański, P., Bingel, J., Diewald, N., Frick, E., Hanl, M., Kupietz, M., Pęzik, P., Schnober, C., & Witt, A. (2013). KorAP: the new corpus analysis platform at IDS Mannheim. *Human language technology challenges for computer science and linguistics. 6th language & technology conference december 7-9, 2013, Poznań, Poland*, 586–587.

Berthold, B. (2023). CREA Working Paper No. 3 Climate Policy Risk and Asset Prices in Switzerland.

Bodmer Mory, F. (2005). COSMAS II-Recherchieren in den Korpora des IDS. *Sprachreport: Informationen und Meinungen zur deutschen Sprache*, *21*(3), 2–5.

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., & Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on language and computation*, *2*, 597–620.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. *Collocations and idioms: Linguistic, lexicographic, and computational aspects*, *23*, 41.

Kupietz, M. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research.

Ort, A., Rohrbach, T., Diviani, N., & Rubinelli, S. (2023). Covering the Crisis: Evolution of Key Topics and Actors in COVID-19 News Coverage in Switzerland. *International Journal of Public Health*, 306.

Purves, R. S., Striedl, P., Kong, I., & Majid, A. (2023). Conceptualizing Landscapes Through Language: The Role of Native Language and Expertise in the Representation of Waterbody Related Terms. *Topics in Cognitive Science*.

Schaber, J., Graën, J., McDonald, D., Mustač, I., Rajović, N., Schneider, G., & Bubenhofer, N. (submitted). The LiRI Corpus Platform.

Vamvas, J., Graën, J., & Sennrich, R. (2023). SwissBERT: The Multilingual Language Model for Switzerland. *arXiv preprint arXiv:2303.13310*.

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., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Dynamically Chaining APIs: from Dracor to TEITOK

Maarten Janssen
Charles University, Faculty of Mathematics and Physics
Prague, Czechia
janssen@ufal.mff.cuni.cz

Abstract

The FAIR principles are meant to ensure that (corpus) data can be reused for other purposes. But reusability is typically only considered from the perspective of static resources, while nowadays, many important data sources are regularly updated. In this paper, we will show how creating one resource from another using the API of both sides can make for a dynamic setup in which two resources can be kept in-sync programmatically, even in the case of changing source data. We will illustrate this by pulling data from the Dracor corpus via the API, and uploading and processing data into a TEITOK corpus also via the API. In the setup in this paper, syncing is done by periodically checking for modified data, but in a more involved integration, an on-update trigger could be used on the source side.

1 Introduction

In the past, it has too often happened that corpora that had been built with considerable effort got lost over time, especially when the server it was hosted on got replaced. Repositories such as GitHub, LINDAT[1], or Nakala[2] are designed to avoid this from happening by providing long-term storage for data. And the FAIR principles[3] are designed to make sure that the data are not merely kept, but can furthermore be reused even if the platform that was originally used to make the data accessible is no longer functional.

FAIR also makes it possible to reuse data more directly after their creation - to provide data that were designed with one purpose in mind for a different target audience, possibly (automatically) enriched in the process. To take a concrete example, it makes it possible to take textual corpora that were collected for any number of purposes (machine translation, speech recognition, manuscript transcription, etc.), enrich them with NLP tools, and make them available as searchable linguistic corpora.

There are numerous example of active reuse in this manner: OPUS (Tiedemann, 2012) actively pursues parallel corpus data, harmonizes them and incorporates the resulting data on their website, both as downloadable data, and as a searchable parallel CWB corpus (Evert & Hardie, 2011). And LINDAT is striving to create a searchable version of any textual data in its repository that are not yet accessible as a searchable corpus and have sufficiently liberal licence - by converting the data to TEI when they are not already in TEI, enrich them with POS, lemmas, and dependency relations using UDPIPE (Straka & Straková, 2017) where needed, and then make the resulting data available as a TEITOK (Janssen, 2016) corpus.

Often, important corpus data are part of an ongoing project, and the repository records are just static versions of data that are still under development. For instance, the Universal Dependencies project consists of a collection of treebanks maintained as Git repositories, and twice yearly the development is halted to perform consistency checks on the data, and create a static release. Also in such cases, the data can be made available in additional tools. In the case of UD, after each release, the data are automatically

[1] https://lindat.mff.cuni.cz/repository/xmlui/
[2] https://www.nakala.fr/
[3] https://www.go-fair.org/fair-principles/

processed, and made available in several searchable environments, including Grew[4] and TEITOK[5]. There are some additional factors to take into account in such cases. There should be no manual corrections in the process, since any manual intervention will be overwritten in the next release. Each new version should be verified changes in the source might make that the conversion scripts no longer work. And the searchable corpus should have a transparent way of indicating which version is being used in the search, and decide whether the prior version should remain accessible. In the case of the TEITOK version, the default search is always in the latest version, but older processed versions remain accessible for the sake of reproducibility. But barring these additional considerations, static versions of "live" corpora can be reused in much the same way as corpora that are fully finished.

But there are live corpora that cannot be (easily) handled by reusing their data for each static version, primarily for two reasons. Firstly, there are resources that simply do not have periodic numbered stable releases that can be used to create a new version of the searchable corpus. This is for instance the case for Dracor (Fischer et al., 2019) or EHRI[6]. And secondly, there are resources that are so extensive that rerunning the entire pipeline periodically requires prohibitive amounts of computational processing, especially where it comes to the NLP pipeline. This is the case for large repositories like Project Gutenberg[7], EEBO[8], or ELTeC (Schöch et al., 2021).

In this paper, we will describe a dynamic setup which can nevertheless create a searchable corpus from such sources by using the API of both resources to first download the data from the source, and then upload them to the searchable target. In the example here, this is done by creating a TEITOK searchable annotated linguistic corpus out of the Dracor corpus.

2 Chaining APIs

In order to create a live searchable corpus out of an under-development source corpus, we will need to establish a list of all the files in the source corpus, convert each file to the format required in the searchable corpus, and then process each converted file. Furthermore, once the initial corpus is established, we will need a mechanism that keeps the searchable corpus in sync with any changes made in the source corpus.

Since there is a Git repository for the Dracor corpus[9], we could simply pull the repository periodically on the target server where the searchable (TEITOK) corpus is located, rely on the git logs to see which files need to be updated after each pull, and then run local conversion and processing scripts to convert the original files to the format required by TEITOK.

However, Dracor is explicitly a programmable corpus, and the API provides the most direct and reliable way to interact with the underlying database. Furthermore, running the conversion scripts directly on the target server forces us to have scripts for the entire conversion pipeline, and it means the scripts have to be run automatically and periodically on the target server, which might be limiting for a more advanced setup (see section 3). So instead, the setup described here relies on the API of Dracor to access to data, and the API of TEITOK to upload and process the files. For source corpora that do not have a dedicated API, but do have a Git repository, git itself can of course be used as an API. The process is very light-weight, since much of the heavy lifting is done by the respective APIs.

2.1 Initial setup

In order to create a searchable corpus in TEITOK, we first need to set up a project for the new corpus, which cannot be done via the API. So via the GUI, we need to provide a name for the corpus, define which parts of the corpus need to become searchable, and ideally provide a landing page text and a graphical template if so desired. And we need to generate an authentication token in order to be able to be able to edit the corpus via the API.

[4]https://universal.grew.fr/
[5]https://lindat.mff.cuni.cz/services/teitok/ud211/
[6]https://www.ehri-project.eu/
[7]https://www.gutenberg.org/
[8]http://eebo.chadwyck.com/home
[9]https://github.com/dracor-org

Once the corpus project is in place, we need to collect all the files that need to be placed in the corpus. In the case of Dracor, the overall corpus consists of a collection of (sub)corpora, so we first need to retrieve a list of the corpora via the API, and then for each corpus, retrieve the list of plays, which can be simply done like this:

```
for corpus in json.loads(requests.get("https://dracor.org/api/corpora").text):
  dramas = json.loads(requests.get(corpus['uri']).text)['dramas']
```

With the list of files to process, we pass each file to a second script that will download the source and upload it to TEITOK. The reason to have a separate script that takes a source file name as argument is that we can call that same script whenever any file needs to be updated later.

For each file, we need to establish a unique ID that will be persistent under updates. And in the case of a multilingual corpus like Dracor, we also need to establish the document languages for NLP purposes. The TEI file rendered by the Dracor API contains an xml:id for each file, which we can use as a unique filename for our target file, and it explicitly specifies the xml:lang for each file except for those in English. The script keeps a local copy of the source file using the unique ID for the filename and proceeds to upload it.

Since the Dracor files are already in a format that the TEITOK API accepts, we can directly upload the file via the API, providing the source file and the file-type, along with the authentication token. There it will in this case simply be copied to into the corpus without conversion. Then, we can ask the API to run the standard NLP pipeline on the file, specifying the corpus language. This will provide inline tokenization to the original TEI file, and if the corpus language is supported by UDPIPE, it will add POS, lemma, and dependency parsing to the file using the REST API of UDPIPE.

Once all the files have been added to our corpus, we can ask the API to create a searchable corpus out of the collection of TEI/XML files, after which we will have an annotated corpus that can be queried online using either CQL or PML-TQ (by default).

2.2 Processing changes

Once an initial version of the corpus has been created, the corpus should be periodically updated to incorporate all the changes in the source corpus. For this, it would be ideal to ask the API to list only those files that were modified since the last update, and process only the files on that list. So if instead of the Dracor API we would be using Git, we could simply make a pull request and then process all the files rendered by the diff: `git diff --name-status HEAD@1..HEAD`

The Dracor API does not currently allow for selecting files by modification date, nor does the list of files per corpus currently provide the last modification date for each file. So the only remedy is to go through the full list of plays in Dracor, and check each one for changes. Since Dracor is of relatively modest size, that is not too problematic. The list we obtain in the same way as in the initial setup, but rather than processing each file, we first check whether the file has been modified. For this, we could in principle check the last revision in the TEI (revisionDesc/listChange/change), but since small updates are not necessarily reflected in the header, we instead verify the file against the copy of the original file kept from the last update. If the new file is identical, we can skip the file, and otherwise we replace the local copy and proceed to process the new version of the file.

By running the update script periodically as a cron job, the Dracor source corpus and the target TEITOK searchable corpus will stay in sync. So at any point, the searchable corpus will have the same textual content as the source corpus, barring any changes made after the last update. So the ideal frequency with which to run the update script depends on the frequency of changes to the source corpus, the overall size of the corpus, and how crucial a perfect match is between the source corpus and the searchable version.

3 Conclusion

In this paper, we have shown that it is possible to reuse corpus data for a different purpose, even in the case of corpora that undergo constant updates, and are either too large to frequently regenerate, or do not have stable versions. And as shown, if the APIs of the two sides (source data and target tool) are properly

set up, reusing is often rather straightforward. This hence highlights why it is important for corpus tools to provide an API, so that their content can be interacted with in a programmatic manner. Or, to put it differently: the addition of an API to a corpus tool significantly increases the reusability of its content.

In the example described in this paper, the update of the searchable corpus is done completely independently of the source corpus. In cases where a searchable corpus is an integral part of the design of the source, a more integrated setup can be designed, in which the script to update a file in the searchable corpus is triggered whenever a file in the source corpus is modified. This can be done either automatically or manually. We are currently looking into such a setup for corpora that are developed in TEI Publisher and for which a searchable version would be highly desirable. This of course is only possible to do from the source server, which is hence one of the reasons to process the files via the API instead of locally, even though in the example setup, no such triggering is performed.

The example used here (Dracor to TEITOK) has its limitations as an example for repurposing corpora via their API. Firstly, Dracor uses the same data format as TEITOK (TEI/XML). But as long as the source data are in a properly machine readable format, and ideally use a standard data format, conversion is often easy: the TEITOK API directly accepts several other formats including some designed for speech or OCR data, and TEI is a well-established format for which numerous conversion scripts are available. Secondly, Dracor does not provide the option to list recent changes. But that merely highlights the need to establish best practice standards for corpus APIs. Thirdly, Dracor is not so large that it becomes impossible to rerun the entire pipeline, which hides the fact that although almost the entire pipeline described here only processed the modified files, except for the last step of generating the CWB corpus from the TEI/XML files, since CQP does not allow for incremental updates, and neither do similar tools like SketchEngine (Kilgarriff et al., 2014). The indexing in CWB is fast enough to be able to rerun that process for larger corpora like EEBO or ParlaMint, but for really large corpora, a corpus tool that can be updated incrementally, based for instance on SQL or SOLR, becomes vital. And fourthly, the example happens to not run across the common issues of legal, ethical, or academic concerns such as copyright, privacy or plagiarism which can plague the proper reuse of data. But these limitations do no affect the fact the inclusion of a properly design API in corpus tools that allows extracting information to repurpose the corpus content has a huge impact on the FAIR-ness of the data.

References

Evert, S., & Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. *Corpus Linguistics 2011*.

Fischer, F., Börner, I., Göbel, M., Hechtl, A., Kittel, C., Milling, C., & Trilcke, P. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. *Proceedings of DH2019: "Complexities", Utrecht, July 9–12, 2019*. https://doi.org/10.5281/zenodo.4284002

Janssen, M. (2016). TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 4037–4043.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, 7–36.

Schöch, C., Patras, R., Erjavec, T., & Santos, D. (2021). Creating the european literary text collection (eltec): Challenges and perspectives. *Modern Languages Open*, *1*. https://doi.org/10.3828/mlo.v0i0.364

Straka, M., & Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. http://www.aclweb.org/anthology/K/K17/K17-3009.pdf

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Analyses of information security standards on data crawled from company web sites using SweClarin resources

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

Subhomoy Bandyopadhyay
Management and Engineering
Linköping University
Linköping, Sweden
subhomoy.bandyopadhyay@liu.se

Svjetlana Pantic Dragisic
Management and Engineering
Linköping University
Linköping, Sweden
svjetlana.pantic.dragisic@liu.se

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

Abstract

With the purpose of analysing Swedish companies' adherence and adoption of the information security standard ISO 27001 and to examine the communicative constitution of preventive innovation in organisations, we have created a dataset of corporate texts from Swedish company websites. The dataset is analysed from multiple interdisciplinary perspectives in close cooperation with organisation researchers and SweClarin researchers using SweClarin tools and resources as well as standard language technology tools. Some analyses require deep reading, which is performed by organisational studies researchers. Initial results have been presented at an organisational studies conference. In this paper, we focus on presenting the research issues, the methods used in the project, and the experience of SweClarin researchers supporting researchers in social sciences. Our contribution is to show how it is possible, through triangulation of human and digital methods, to increase the credibility and validity of a digitally acquired data set and subsequent research findings. In our view, a combination of human deep reading (organisation researchers), contextual dictionary verification (organisation and management studies) and language technology (sentiment analysis) can help to sensitise computational text analysis for medium-sized data sets.

1 Introduction

Preventive innovation differs from ordinary innovation. The innovation literature claims that the economic benefits of preventive innovation to organisations, for instance, for avoiding environmental pollution, protecting human health or ensuring information security, are mainly intangible, often time-delayed and adopted for incidents that may never occur (Rogers, 1995). To address these challenges, organisational communication seems to be crucial to increase the potential of economic recognition for preventive innovation.

Therefore, we draw on the "communicative constitution of organisations" view to explore how preventive innovations are communicatively constituted. Using the example of the information security standard ISO/IEC 27001, we examine how communication of preventive innovations is shaped by its adopting organisations. We analyse texts about the information security standard ISO/IEC 27001 on Swedish corporate websites supported by computational tools for web scraping and language analyses. As a result, we first identify three communicative practices of data governance termed agency, stewardship and brokerage, and second, provide evidence that organisations' communication also depends on whether they receive direct or indirect economic recognition for their preventive innovation.

We contribute a meaningful combination of deep reading of humans (researchers), dictionary verification for a specific context (innovation research) and language technology (sentiment analysis) to a meaning-centred and situational understanding of preventive innovation. Our analysis enhances Rogers' perspective by challenging the classification of preventive innovations as mere "isolated, static objects or practices", unveiling their dynamic interplay with organisational members — simultaneously influencing and being influenced — i.e., are enacted communicatively by organisations. Contrary to Rogers' assumption, we also provide initial evidence that preventive innovations can very well achieve economic recognition by constituting different meanings of preventive innovation.

This paper will focus on the methodology, rather than delving into the theoretical underpinnings. We illustrate the potential of SweClarin and language technology analyses for investigating organisational communication and the production of meaning in their texts.

2 Research design

Using ISO/IEC 27001 as an example to study the communication of preventive innovations, our research design followed three steps, see Figure 1. We first generated a dataset of Swedish corporate websites of all sectors and scraped the content for ISO/IEC 27001 related paragraphs of the text corpus. Second, we categorised the identified companies manually according to their adoption (of preventive innovation) approach. Finally, we conducted analyses on the language used in the paragraphs relating to ISO/IEC 27001 on these websites.

Regarding the first step, as a complete dataset of all websites of Swedish companies does not exist as open access, we contacted several institutions to retrieve this data. We approached Sweden's company registration office, Bolagsverket, and Statistics Sweden (SCB) to get access to company names, identification numbers, sector affiliations and innovation indicators. However, Bolagsverket and SCB could not provide a database with company URLs. We, therefore, analysed 400 company names on Nasdaq Nordic (https://www.nasdaqomxnordic.com/) through scripts that generate web addresses in order to understand how company URLs can be constructed, and used that to generate 120 million possible URLs from the 2.4 million registered companies listed on Bolagsverket. These URLs were then tested to check how many of them were actual websites. These websites were then scraped in September 2020. We scraped up to 50 connected web pages of each site to grasp sufficient content (cf., Kinne and Lenz (2019)). Out of all scraped websites, we found 472 which contained the filter phrases 'ISO 27001', 'IEC 27001', 'IEC 270' or 'ISO 270'[1].

After we had identified the 472 websites[2], as a second step, we manually analysed each company's website by visiting their URLs to verify the scrapped data. This hands-on scrutiny of the corporate websites aimed to refine the extracted information regarding companies' certifications, business sectors, models, and value propositions. After removing duplicates and further non-Swedish companies in the dataset, we were left with 353 websites of Swedish companies. We categorised these companies according to the criteria 'certified' or 'non-certified', following a suggestion by Mirtsch et al. (2020). Their findings reveal that a third of the companies that adopt ISO/IEC 27001 do so through certification, with the remainder opting for non-certified pathways. Furthermore, our findings revealed a variety of companies: some integrated ISO/IEC 27001 consulting or training into their business models, while others, lacking certification and refraining from offering consulting or training services, solely referenced certified clients, customers, and suppliers on their websites. Based on this initial categorisation, we identified six distinct types of preventive innovation adoption, denoted as 11, 12, 21, 22, 31, and 32.[3] into which each company belongs.

In addition, two text corpora were generated from all identified company websites, one in Swedish and one in English. We use fastText (Joulin et al., 2016) to separate the sentences. For each company, we take each sentence and place it in an English or a Swedish text file, i.e. a company can have two files, one

[1] Including variants such as iso-27001 and Iso 270.

[2] Available at https://www.ida.liu.se/~arnjo82/472_webpages

[3] The first digit (1, 2, or 3) denotes the three data governance approaches: Agents, Stewards, and Brokers, whereas the second digit (1 or 2) signifies (in)direct economic benefits resulting from preventive communication adoption, evaluated based on ISO/IEC 27001 training/consultation provision.

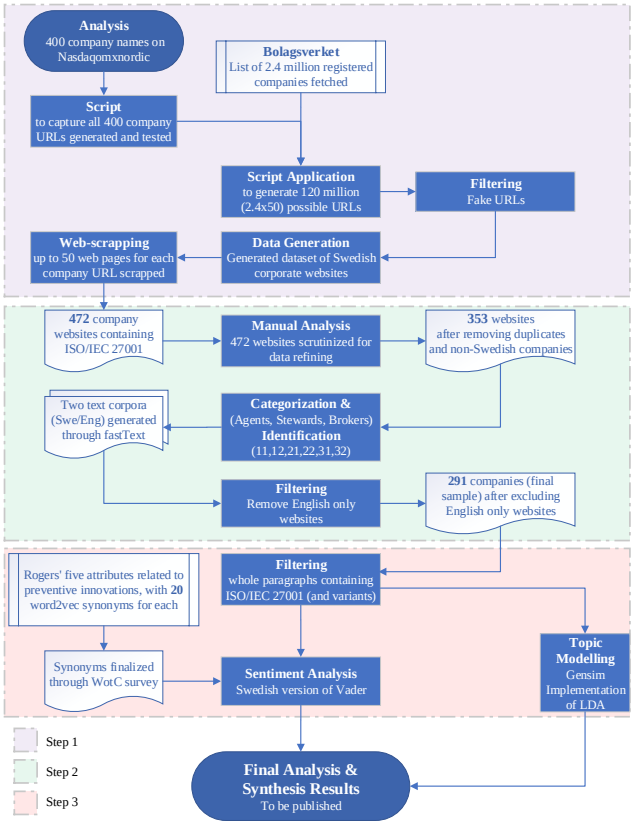


Figure 1: Overview of the process.

with English text and one with Swedish. The English text corpus, spanning around 450 pages, underwent manual analysis through deep reading, revealing that over 50% of the dataset consisted of inconsequential noise such as ads, menus, contact details, and website cookies. As an outcome of this analysis and in pursuit of methodological rigour through Swedish sense-based sentiment analysis (elaborated upon below), companies with English only websites were excluded from the sample, resulting in 291 companies (final sample size)[4] with websites in either Swedish or both Swedish and English. Although certain Swedish websites maintained English versions, it is noteworthy that, for analytical efficiency, the term ”Swedish only” pertains solely to these 291 entities, since the English text corpus had been excluded from further analysis. Table 1 depicts the number of sentences and words for each adoption type for the 291 companies with Swedish only text on their web pages.

This resulted in a text corpus of close to 9 million words, see Table 1. Content analysis on these texts was performed to demonstrate how preventive innovation is manifested within the communication of the six identified adoption approaches. To aid this analysis word clouds were created using the WordCloud package[5] and topic analysis using the Gensim implementation of Latent Dirichlet Allocation (LDA) (Blei

[4] Available at https://www.ida.liu.se/~arnjo82/291_filtered_webpages
[5] https://pypi.org/project/wordcloud

Adoption type	11	12	21	22	31	32
Number of companies	103	10	81	41	19	37
Number of sentences	197.131	11.225	127.880	29.404	20.462	82.390
Number of words	3.374.348	187.630	2.133.516	547.543	351.837	1.683.044
ISO paragraphs	520	88	401	133	17	372
Sentences in ISO paragraphs	8356	1153	4404	2248	561	38817

Table 1: Descriptive statistics for the Swedish companies in each adoption type

et al., 2003). We also translate all Swedish texts to English using googletrans[6], as not all organisational studies researchers are fluent in Swedish.

To assess the relevance and usefulness of preventive innovations along five attributes (as suggested by Rogers), we use sentiment analysis. We want to compare the overall sentiment for each attribute and also compare the sentiment when ISO/IEC 27001 is presented. The paragraphs in the files containing 'ISO/IEC 27001', and its possible variants, were filtered out of each text to be used for sentiment analysis. We use the context in which an ISO/IEC 27001 sentence occurs, i.e. the whole paragraph, as it is scraped from the web. This filtering resulted in a considerably smaller number of paragraphs and the sentences within them (Table 1).

We use sentiment analysis along five attributes: relative advantage, compatibility, complexity, trialability and observability (Rogers, 1995). To capture various uses of the attributes, synonyms were generated for each attribute by using the Gensim package Řehůřek and Sojka (2010). For each attribute we generated 20 synonyms using seeds, in Swedish, that reflected the various attributes. For three of the attributes, we generated a second set of synonyms using different seeds. The general applicability of these twenty computer-generated synonyms in the Swedish colloquial language was assessed through a wisdom-of-the-crowd (WotC) survey approach (Surowiecki, 2004). An online Microsoft Forms survey with these twenty synonyms was sent to native Swedish speaking innovation and entrepreneurship researchers at Linköping University to compile a final set of synonyms for the five attributes.

For sentiment analysis, we use a Swedish version of Vader (Hutton & Gilbert, 2014) that considers a word's sense. Vader is a lexicon and rule-based sentiment analyser. The lexicon in English Vader comprises 5500 lexical entries with sentiment scores between +5 and -5. We used the Swedish SenSALDO 0.2 sentiment lexicon (Rouces et al., 2019) with sentiment scores -1, 0 and +1. SenSALDO 0.2 comprises 12287 lexical entries of which 8893 are unique words. Word sense disambiguation with the SenSALDO 0.2 lexicon is achieved by first parsing the texts using the Sparv pipeline[7] (Borin et al., 2016). Vader also uses booster words, such as amazingly, to further refine the sentiment analysis. The booster dictionary used in these analyses is a enhanced version of the Swedish dictionary used for sentiment analysis of e-mail conversations (Borg & Boldt, 2020) and comprises 89 items. The version used in this project, using the SweClarin SenSaldo resources, has also been used in a project on analysing Swedish official texts (Ahrenberg et al., 2022).

Data from websites are very noisy containing repetitions, menu items, contact information, adverts, etc that need to be handled. Standard crawling packages provide some cleaning of the texts but there is still much that is, for instance, not syntactically correct. Despite this, we find that the SweClarin Sparv pipeline is robust and provides an analysis that can be used by the sentiment analyser.

3 Conclusions

In this study, we started with the idea of reviving the concept of preventive innovation given the attention this type of innovation is receiving nowadays. We have chosen to explore the adherence and adoption of

[6]https://pypi.org/project/googletrans

[7]https://spraakbanken.gu.se/sparv/#/sparv-pipeline

the ISO/IEC 27001 information security standard as an example of preventive innovation addressing cybersecurity risks as one of the great challenges of our time. Using web scraping tools and computational linguistics (and content analysis on top of that), we were able to extract and analyse large amounts of text. These texts on preventive innovation ISO/IEC 27001 include communicative efforts published on the websites of companies operating in Sweden, telling us about the way these companies are adopting the standard. We have identified different adoption approaches and related modes of data governance. These results also help us understand that the original concept as introduced by Rogers (1995) needs to be improved in terms of opportunities to derive economic benefits from preventive innovation. By relating the adoption approaches to the different modes of data it could be shown that a meaningful adoption of preventive innovations can already take place at an early stage.

The close cooperation between the organisational studies researchers and the SweClarin language technology researchers has been imperative for the success of this project. Based on the needs of the organisational studies researchers' various analyses have been performed, and assessed. It was, for instance, initially assumed to be important to use topic models to guide the deep readings. The topic maps, however, turned out to be rather diverse and did not form a clear picture of the various adoption types. Instead, word clouds were developed that gave a better, but not sufficient, analysis of the data. In further discussions with the organisational studies researchers, we decided to try sentiment analysis, which turned out to give useful results on its own as well as the possibility to generate quotes from the texts for each sentiment ranked by its score. This gave organisation researchers the opportunity to see a quantification of the meanings to further aid the deep readings.

References

Ahrenberg, L., Holmer, D., Holmlid, S., & Jönsson, A. (2022). Analysing changes in official use of the design concept using sweclarin resources. *Proceedings of the 2022 CLARIN Annual Conference*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Borg, A., & Boldt, M. (2020). *Using vader sentiment and svm for predicting customer response sentiment, expert systems with applications* (Vol. 162). Elsevier.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

Hutton, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kinne, J., & Lenz, D. (2019). *Predicting innovative firms using web mining and deep learning* (tech. rep.). ZEW Centre for European Economic Research, Discussion Paper. 01/2019 (19-001). http://ftp.zew.de/pub/zew-docs/dp/dp19001.pdf

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

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora [http://is.muni.cz/publication/884893/en]. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rogers, E. M. (1995). *Diffusion of innovations*. The Free Press.

Rouces, J., Tahmasebi, N., Borin, L., & Eide, S. R. (2019). Sensaldo: Creating a sentiment lexicon for Swedish. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 4192–4198.

Surowiecki, J. (2004). *The wisdom of crowds*. NY, NY: Anchor.

Linguistic Resources and Tools for Ukrainian: Grounds for Creating a K-center

Olha Kanishcheva
University of Jena, Germany
SET University, Ukraine
kanichshevaolga@gmail.com

Maria Shvedova
University of Jena, Germany
mariia.o.shvedova@lpnu.ua

Abstract

This paper provides an overview of open resources and instruments for Natural Language Processing (NLP) in Ukrainian, experience in Ukrainian Natural Language Processing, and shows the grounds for the creation Ukrainian CLARIN K-center. The authors discuss different resources, tools and projects for Ukrainian NLP, and provide an overview of corpora for Ukrainian, including the *General Regionally Annotated Corpus of Ukrainian (GRAC)*, the *Corpus Project of the Laboratory of Ukrainian*, the *Lang-uk corpus project*, among others. The authors argue that creating a K-center for the Ukrainian language will contribute to the transfer of experience of Ukrainian NLP technologies, sharing knowledge with other European centers, and faster integration of Ukrainian resources into the EU network.

1 Introduction

Every year, Ukrainian scientists, developers, and teachers become increasingly involved in European projects and grants in order to adopt the best European practices. In 2022, the Agreement on Ukraine's participation in the EU programs "Horizon Europe" and "Euratom" was ratified which ensured the access of Ukrainian scientists and innovators to the financial resources of the programs and, in the long term, will contribute to the active integration of Ukraine into the European Research Area. However, this integration is quite slow so far and people do not receive the necessary knowledge comprehensively.

The situation is especially difficult in the Humanities and Social Sciences (SSH). This is where the K-centers can help to fill in numerous gaps. First, the K-centres of CLARIN are designed to be user-friendly and accessible to researchers without specialized technical expertise and is intended to support a wide range of research in the humanities and social sciences. Second, a K-centre typically includes resources and tools such as annotated corpora, lexical databases, and software for text analysis and visualization. These resources and tools are designed to help researchers analyze and understand language data in various languages, and to support research on topics such as language variation, language change, and language acquisition. Thus, we propose to create a K-center for the Ukrainian language, which will allow the integration of existing resources into the CLARIN infrastructure and also the creation of new ones based on certain metadata requirements, etc. The same holds for Ukrainian language processing tools.

2 Overview of open resources and tools for Ukrainian NLP

2.1 Morphological dictionary of the Ukrainian language and main corpora

Over the past 30 years, a fairly large number of different corpora have been created for the Ukrainian language. There is still a lack of specialized data. Below, information about the most important resources is given; a more comprehensive list of major open Ukrainian corpora can be found in Table 1.

The Large Electronic Dictionary of Ukrainian (VESUM)[1] counts more than 416 thousand lemmas and is being constantly updated. It contains information on the inflection of the words; non-standard and alternative word forms and their alternatives are highlighted; abbreviations and contractions accounted

[1] https://github.com/brown-uk/dict_uk

for; information on some alternative orthographic norms is included; it encompasses a large database on proper names; it is synchronized with the Ukrainian gazetteer, including place names that appeared after the decommunization; it features a handy system of marking inflectional types and tags that enables easy updates and regrouping of existing words; it also contains data on some rare and spoken forms (Starko & Rysin, 2020). In 2022 VESUM was automatically converted to the UniMorph scheme: https://github.com/unimorph/ukr.

General Regionally Annotated Corpus of Ukrainian (GRAC) [2] is the largest manually compiled reference corpus of Ukrainian (Shvedova, 2020). The size of the latest version (GRAC.v.16, 2023) is 1.9 billion tokens. It includes texts from 1816 to 2023 written both in Ukraine and in the diaspora, coming from printed, recorded, and handwritten sources. It features fiction, journalism, academic texts, laws, religious literature, letters and diaries, and transcripts of oral speech. The texts in GRAC are annotated by the author, date of creation and publication, style, genre, and region. GRAC is morphologically annotated, and the lemmatization is carried out on the basis of VESUM. The morphological analysis system has some special rule-based tools for processing non-standard spellings applied to the historical part of the corpus.

The **Corpus Project of the Laboratory of Ukrainian** [3] contains several corpora and a dedicated morphological analyzer. The corpora include a treebank with manual disambiguation and manual tagging (140 thousand tokens), a web corpus "Zvidusil" with automatic syntactic annotation (about 3 billion tokens), parallel corpora (Kotsyba & Moskalevskyi, 2018).

The **Lang-uk corpus project**[4] provides collections of Ukrainian online press, fiction, and Wikipedia available for download, totaling 665 million tokens (UberText corpus (Chaplynskyi, 2023)), a corpus of law and legal acts counting 579 million tokens, a corpus annotated for named entities and also a build-up model for automatic annotation of named entities (people, organizations, locations, and others); different gazetteers, simple tokenizer (splitting text into paragraphs, sentences, and tokens), vector models trained on different corpora.

A corpus of Ukrainian parliamentary transcripts, compiled according to the Parla-CLARIN TEI schema for corpora of parliamentary proceedings as part of the **ParlaMint project**[5]. The same source transcripts were also used to create a code-switching corpus for the study of Ukrainian-Russian bilingualism (Kanishcheva et al., 2023).

2.2 The tools of processing the Ukrainian language

The field of Natural Language Processing (NLP) has seen remarkable growth in the development of tools specifically tailored for the Ukrainian language. Below there are some of them:

Nlp-uk — an instrument based on the VESUM dictionary and the LanguageTool engine. Supports tokenization, lemmatization, POS analysis, and basic disambiguation.

Pymorphy2 — a morphological analyzer without disambiguation; the Ukrainian language is supported via the old version of VESUM.

Stanza — the Stanford library for language processing; it supports Ukrainian using the UD corpus. Features models for tokenization, lemmatization, POS and syntactic analysis.

LanguageTool — spelling, stylistic, and grammar checker, which helps to correct and paraphrase texts.

Stemmer for Ukrainian language - a new stemmer for the Ukrainian language (tree_stem) created via machine learning.

These examples highlight the growing ecosystem of NLP tools for the Ukrainian, fostering improved communication, data analysis, and content generation capabilities. As these tools continue to evolve, they offer promising opportunities for further innovation and language empowerment. More information about tools and corpora is presented in this link https://github.com/asivokon/awesome-ukrainian-nlp

[2] http://uacorpus.org
[3] https://mova.institute
[4] http://lang.org.ua/uk/corpora
[5] https://www.clarin.eu/parlamint

3 Integration of Ukrainian language resources and analysis tools into CLARIN

Integration of Ukrainian resources and tools (LRT) with CLARIN refers to the process of making LRT available through the CLARIN infrastructure. It allows researchers to access and use these resources and tools more easily. Integration into the CLARIN infrastructure involves several steps, including:

- Preparation of language resources and tools: Language resources and tools should be prepared in a format that is compatible with the CLARIN infrastructure. This may involve converting data to a specific format or developing software that can be accessed through the CLARIN portal.
- Metadata creation: Metadata, or descriptive information about the language resources and tools, have to be created in a standardized format that can be used to search and retrieve these resources through the Virtual Language Observatory.
- Quality assurance and curation: Language resources and tools are subject to quality assurance and curation by the CLARIN centers, to ensure that they meet the standards for inclusion in the infrastructure.

Accordingly, once language resources and tools have been integrated with CLARIN, they are made available to researchers through the CLARIN portal, where they can be searched and accessed using a variety of tools and interfaces. Integration with CLARIN provides researchers with access to a wide range of high-quality language resources and tools and helps to facilitate collaboration and sharing of data across disciplines and institutions.

4 Conclusions

In this paper, we proposed the idea to create a K-center for the Ukrainian language and proposed to unite both researchers in SSH and Natural Language Processing engineers. The creation of such a center will allow to adapt many existing tools and methods for processing the Ukrainian language. The integration with CLARIN will provide easier access to language resources and tools for researchers in the humanities and social sciences. Moreover, such a center will provide consultations on issues related to the processing of text in the Ukrainian language, on LRT, and conduct training seminars and consultations. It is planned to inaugurate such a center in September-October 2023.

References

Chaplynskyi, D. (2023). Introducing UberText 2.0: A corpus of modern Ukrainian at scale. *Proceedings of the Second Ukrainian Natural Language Processing Workshop*, 1–10. https://aclanthology.org/2023.unlp-1.1

Kanishcheva, O., Shvedova, M., Kovalova, T., & von Waldenfels, R. (2023). The parliamentary code-switching corpus: Bilingualism in the Ukrainian parliament in the 1990s-2020s. *Proceedings of the Second Ukrainian Natural Language Processing Workshop*.

Kotsyba, N., & Moskalevskyi, B. (2018). An essential infrastructure of Ukrainian language resources and its possible applications. *SlaviCorp 2018. Book of Abstracts*, 94–95. https://slavicorp.ff.cuni.cz/wp-content/uploads/sites/144/2018/09/SlaviCorp2018_Book_of_Abstracts.pdf

Shvedova, M. (2020). The general regionally annotated corpus of ukrainian (grac, uacorpus.org): Architecture and functionality. *CEUR Workshop Proceedings. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2020)*, *Volume I: Main Conference*, 489–506. http://ceur-ws.org/Vol-2604/paper36.pdf

Starko, V., & Rysin, A. (2020). Velykyj elektronnyj slovnyk ukraïns'koï movy (VESUM) jak zasib NLP dlja ukraïnsḱoï movy. *Halaktyka Slova. Halyni Makarivni Hnatjuk*, 135–141.

Corpus	Size	Texts included	Access
http://uacorpus.orgGRAC	1.9 bln tokens	Various texts (reference corpus)	Searchable online
Corpus of Ukrainian (Taras Shevchenko National University of Kyiv)	120 mln tokens	Journalism, fiction, academic, legal, poetic	Searchable online, user interface in Ukrainian
Ukrainian Brown corpus	462,000 tokens	Balanced, manually annotated corpus	Available for download
Ukrainian Treebank (Laboratory of Ukrainian)	140 thousand tokens	Different genres	Searchable online, available for download
Zvidusil: a web corpus with syntactic annotation (Laboratory of Ukrainian)	3 bln tokens	Web texts (2018)	Searchable online
Ukrainian Web Corpus (Leipzig University)	1,5 bln tokens	Web texts (2014)	Searchable online
Web Corpus Araneum Ucrainicum	125 mln tokens (“Minus”) and 1,25 bln tokens (“Maius”)	Web texts (2014, 2015, 2021, 2022)	Searchable online, registration is required
Polish Automatic Web corpus of Ukrainian language (PAWUK)	700+ mln tokens	Web texts (news sites, telegram, twitter, YouTube), downloaded daily from March 2022	Searchable online
Lang-uk	600 mln tokens	News, Wikipedia, fiction, web	Available for download
Ukrainian corpus of the Chtyvo library	600 mln tokens	Books: fiction, academic texts, journalism	The search is exact (without lemmatizing, morphology or correcting mistakes) and available online
Parallel with English, Polish, French, German, Spanish, Portuguese (Laboratory of Ukrainian)	5 mln tokens	Fiction	Searchable online
Parallel with Russian (Russian National Corpus)	9 mln tokens	Fiction, journalism	Searchable online
UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language (Grammarly)	34,000 sentences	Texts with errors	Available for download

Table 1: Open Ukrainian Corpora.

Korpusnik: a corpus summarizing tool for Slovene

Iztok Kosem
Jožef Stefan Institute &
Faculty of Arts, University of Ljubljana
Ljubljana, Slovenia
iztok.kosem@ijs.si

Jaka Čibej
Jožef Stefan Institute &
Faculty of Arts, University of Ljubljana
Ljubljana, Slovenia
jaka.cibej@ijs.si

Kaja Dobrovoljc
Jožef Stefan Institute &
Faculty of Arts, University of Ljubljana
Ljubljana, Slovenia
kaja.dobrovoljc@ijs.si

Simon Krek
Jožef Stefan Institute &
Faculty of Arts, University of Ljubljana
Ljubljana, Slovenia
simon.krek@ijs.si

Abstract

In this paper, we present Korpusnik, a corpus summarizing tool for Slovene, which is being developed as part of the CLARIN.SI infrastructure. The tool offers a simple and clear overview of the most relevant information (e.g., collocations, example sentences, distribution by text type, year of publication, and source) from five corpora of Slovene: the Gigafida Corpus of Written Standard Slovene, the Gos Corpus of Spoken Slovene, the Trendi monitor corpus of Slovene, the JANES Corpus of Internet Slovene, and the OSS Corpus of Slovene Scientific Texts. Special attention in the design of the tool has been paid to accessibility, especially for people with disabilities.

1 Introduction

CLARIN.SI is the Slovene national node of the European CLARIN infrastructure, with the main aim of supporting researchers and other interested parties in their research in the use and production of language data. In addition to expert and technical support, CLARIN.SI provides services such as a repository for storing language resources and tools (currently containing over 150 datasets or tools), online concordancers (with access to nearly 100 corpora), and other tools and services (e.g., for annotation). An important aspect of CLARIN.SI is also raising awareness of its activities at the national and international levels, which is done in the form of presentations, training events, blogs, etc.

With such a vast array of resources and services, one of the challenges of CLARIN.SI has been in how to increase its user base. On the one hand, potential users who may be less proficient in the use of concordancers and related tools can be deterred by the required knowledge level of the KonText and NoSketch Engine tools. Also, these concordancers do not offer the option to compare information across different corpora, which might interest not only researchers but also other users in the language field (e.g. translators, teachers) or even the general public.

Recently, an opportunity presented itself to address this problem. In 2022, the Jožef Stefan Institute as a member of CLARIN.SI successfully obtained funding from the Slovenian Ministry of Culture for a project called Upgrading the CLARIN.SI portal: Corpus summarizer and text analyzer (SLOKIT),[1]

[1] https://slokit.ijs.si/

which aims to develop two tools: a corpus data summarizing tool and a text analyser. The project will result in two other deliverables: the improved versions of the Gigafida Corpus of Written Standard Slovene and the Gos corpus of Spoken Slovene. The new version of Gigafida will fix an issue with text segmentation for newspaper texts, as the majority of newspaper texts in the current version consist of entire daily issues instead of individual articles. The Gos corpus will be morphosyntactically annotated with better models and improved by aligning speech transcriptions with sound recordings at the level of individual tokens; in addition, an automatic approach to assigning the most typical pronunciation for tokens will be tested. In our presentation, we focus mainly on the corpus summarizer. An important part of the project is also the interface design, as one of the specific requirements of the project call was to ensure that the tools are accessible to users with disabilities, i.e. hearing-impaired, sight-impaired, or movement-impaired.

2 Related tools

At the time of writing this paper, only a handful of corpus summarizing tools were publicly available. Two online tools that we consulted extensively for our project purposes were Word at a Glance (Machálek 2020), hosted by the Institute of the Czech National Corpus, and the BNClab (Brezina et al. 2018). Word at a Glance provides information (profiles) on Czech words, drawing on the written and spoken corpora of Czech, and also offers the possibility to compare the profiles of two words and to search for translations of queried words into other languages. The information is presented on a single page consisting of several sections the user can open or close based on their interests. The tool includes various forms of data presentation, from text to diagrams, a word cloud, and a map display for spoken data. Similar features can be found in the BNClab, with the main difference being that the BNClab draws the data from two versions (1994 and 2014) of a single corpus, namely the British National Corpus (BNC). It contains separate sections on usage (concordances), change (comparing the data from 1994 and 2014 versions), gender, age, social class, region (in all four categories of the authors), and the comparison of the use in speech and in writing.

Even more advanced corpus tools offer only limited data visualization features. For example, the recent version of the Sketch Engine tool (Kilgarriff et al. 2004) and its free version noSketch Engine offer diagrams and graphs for the text type analysis of the entire corpus, and for the distribution of the queried hits in the corpus. Some visual elements are also used by AntConc (Anthony 2022), NOOJ (Silberztein 2003-), Voyant Tools[2] and others. Nonetheless, these graphical elements are normally used only to offer an alternative to complex data visualisation, rather as the main exploratory features.

Overall, we can conclude that while there have been attempts made to bring corpus data closer to researchers who are not experts or regular users of corpus tools, they still come with fairly steep learning curves in terms of how corpus tools work, which information they contain, and how to interpret them. Furthermore, different possibilities of visualisation have still not been fully exploited.

3 Korpusnik – a corpus summarizing tool

We approached the development of Korpusnik with several goals in mind. Firstly, we wanted to provide users of CLARIN.SI infrastructure with a quick overview of the use of words in five different language corpora of Slovene. Secondly, we wanted to include features that would demonstrate the wide variety of data from different corpora in a user-friendly and easy-to-understand manner. Thirdly, we wanted to ensure that the design of Korpusnik is developed according to the guidelines for websites for different groups of users with disabilities.

Our very first activities in the development of Korpusnik included focus groups with potential users, especially people with disabilities. As most of them were not familiar with corpus tools and corpora, we initially asked them to comment on the usefulness of the Word at a Glance website (as the best approximation to what we were developing), and later, after the conceptualisation of Korpusnik, also asked for their comments about the planned features. Web Content Accessibility Guidelines[3] were listed as a useful reference document when preparing any website that targets users with disabilities. The comments

[2] https://voyant-tools.org/
[3] https://www.w3.org/TR/WCAG21/

and recommendations of the participants in our studies could be divided into three categories: visual elements (including fonts etc.), interactivity, and sustainability. It was recommended that clear titles for different sections or diagrams should be provided, a suitable font used (at least 14pt), and suitable contrast (avoiding the combination of white and blue, or yellow and green). In terms of interactivity, the often-expressed view was that the use of a keyboard for all the functionalities should be supported. The use of technologies such as speech synthesis was considered very useful. It was suggested that as for Slovene more corpora will be used, it would be useful to not show all the information on one page and to consider providing some summary in a text format rather than relying purely on graphical elements. As far as sustainability goes, it was suggested that we should use a standard programming language, which would facilitate any updates to the tools used by the users, or to the programs used by Korpusnik (e.g. Java, Python).

The first step was to select the corpora from which we would extract the necessary information. One of the conditions was that the corpora are available in the CLARIN.SI infrastructure, more specifically in one of CLARIN.SI concordancers. Moreover, we wanted the selected corpora to represent different aspects of the Slovene language. The following corpora were selected:

- Gigafida (the current version is 2.0) is a reference corpus of standard written Slovene containing over 1.3 billion tokens. It is comprised of daily news, magazines, a selection of web texts (a certain portion of which covers news texts as well), and different types of publications (fiction, school books, and non-fiction). The current version covers the period from 1991 to 2018.
- Trendi (the current version is 2023-02) is a monitor corpus of Slovene and contains just over 700 million tokens. It comprises news from 107 Slovene media websites published by 72 different publishers. Trendi 2023-02 covers the period from January 2019 to February 2023, complementing the Gigafida 2.0 reference corpus of written Slovene.
- Gos (the current version is 2.0) is a reference corpus of Slovene speech, containing 2.5 million tokens. The current version includes about 300 hours of speech, 127 thousand utterances, and 1,500 texts.
- OSS (the current version is 1.0) is a corpus of scientific writing in the Slovenian language, containing over 3.2 billion tokens. The texts were gathered from the Open Science Slovenia portal (https://openscience.si). The OSS corpus consists of over 150 thousand monographs, articles, bachelor's, master's and doctoral theses, advanced textbooks, reviews and similar texts, mostly published between 2000 and 2022 by Slovenian universities, research institutions, etc.
- JANES (the current version is 1.0) is a corpus of Internet Slovene containing over 250 million tokens. It comprises texts from blog sites, forums, comments of news, tweets, and wiki sites.

For each corpus, the types of information considered most relevant for the users were selected, and it included relative frequency, list of word forms, collocations, and distribution of hits according to different criteria depending on the corpus (text type, year or month of publication, domain, publisher, gender of the author, region of the author(s), sentiment, etc.), and examples of use.

The main principle when developing the backend of Korpusnik was to use the API calls to obtain data rather than design a database with stored queries. This decision was made because we could not anticipate all possible queries and we did not want to limit user searches, and also because using the API enables us to quickly switch to new versions of the corpora once they become available. Out of the three concordancers used by CLARIN.SI – KonText, noSketch Engine (Crystal), noSketch Engine (Bonito) – noSketch Engine (Bonito) proved to be the most suitable as it supports the JSON format of output which significantly speeds up the calls.[4]

In terms of graphical design, we are collaborating with a design company to create an aesthetically pleasing, easy-to-use, and not overwhelming online tool. In order to achieve these characteristics, we decided to take a different approach to structuring the webpage as used for example by Word at a Glance. The website offers two types of queries, a single-word search and a comparison search. In the single-word search, one is provided with an overview of the main highlights of the word's use in the five corpora (the Highlights tab). These highlights include a comparison of the word's relative frequency in

[4] NoSketch Engine (Crystal), which also supports JSON, was causing some issues with the JSON output.

the five corpora, word usage in the period covered by the reference written and monitor corpora, collocations from the three corpora (Gigafida, Trendi and OSS), and examples of use from each corpus. In addition, the user can also examine the use of the word in each of the five corpora in separate tabs, with much more detail being provided (e.g. for the GOS corpus of Slovene speech, one can see the regional distribution of the word, distribution by gender, age and education of the speaker etc. The comparison page is intended to compare the uses of two words, using similar graphic solutions as in the individual mode, but in a parallel manner (Figure 2).

A key feature found on all the pages of Korpusnik is a short summary of the main observations about the word's use, called Main points ('Glavne točke'; see Figure 1). This summary is available at every subpage, i.e., Highlights and at the summary page of every corpus. This feature is a direct result of our surveys and discussions with focus groups. The summaries are presented using a pre-prepared template sentences, and a set of parameters which trigger the inclusion of a certain template.

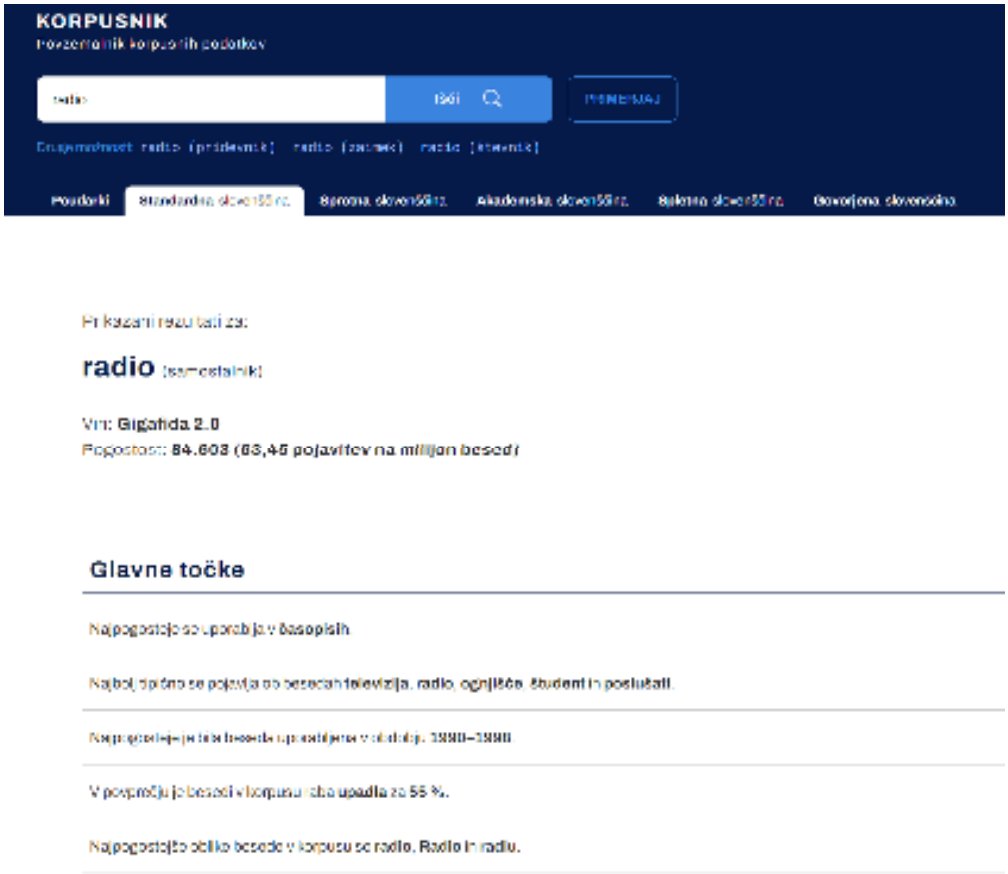


Figure 1: Main points for the noun *radio* in the reference written corpus Gigafida 2.0.

One of the important decisions was related to the type of search used by the tool. We did not want to include any advanced options such as CQL, simply because those are for advanced users and should be done in one of the concordancers. The default search is a search by lemma. The first call is about all the POS tags, and out of them, the information for the most frequent lempos is immediately shown to the user, with others available under the search box as alternatives. The reason for choosing this approach lies in the fact that there could be problems in automatic annotation (e.g. in Trendi *covid* is found as a noun, an adjective, an adverb, an interjection etc.) or in the overlap of forms belonging to different lemmas (*lev* can be a noun, i.e. lion, but also an adjective, i.e. left).

Figure 2: A comparison of the use of nouns *koronavirus* and *covid* from 1990 to 2023 (part of the Highlights tab).

4 Conclusion

At the time of writing, the development and design of Korpusnik was in its final stages. At the conference, a demo of the tool will be given, along with a closer examination of the backend, including an API for anyone wanting to use the information summarized in Korpusnik for other purposes. A part of the project is also dissemination activities, including workshops and other events, social media postings, and mailing list notifications.

We hope that Korpusnik will attract more researchers and other users who have so far not used corpora and other parts of the CLARIN.SI infrastructure or have used it rarely. Moreover, we also hope that other CLARIN ERIC members will recognize the potential of such tools and perhaps include it in their own infrastructures.

References

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Brezina, V., Gablasova, D. & Reichelt, S. (2018). BNClab. http://corpora.lancs.ac.uk/bnclab [electronic resource], Lancaster University.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the 11th Euralex International Congress*. Lorient: Universite de Bretagne-Sud, 105-116.

Machálek, T. (2020). Word at a Glance: Modular Word Profile Aggregator. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, 7009–7014.

Silberztein, M. (2003-). NooJ Manual. Available for download at: http://www.nooj4nlp.org

Sharing the Finnish Dark Web Marketplace Corpus (FINDarC)

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

Teemu Ruokolainen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
teemu.ruokolainen@tuni.fi

Lasse Hämäläinen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
lasse.hamalainen@tuni.fi

Tuomas Harviainen
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
tuomas.harviainen@tuni.fi

Abstract

We discuss the archiving procedure of a corpus comprising posts submitted to Torilauta, a Finnish dark web marketplace website. The site was active from 2017 to 2021 and during this time one of the most prominent online illegal narcotics markets in Finland. As a result of the presented work, a reduced version of the corpus, Finnish Dark Web Marketplace Corpus (FINDarC), has been archived in the Language Bank of Finland. Researchers can apply for access rights to the corpus under the CLARIN RES licence.

1 Introduction

Torilauta was a dark web marketplace website. The site was active from 2017 to 2021 and during this time one of the most prominent online illegal narcotics markets in Finland. Functionally, the site consisted of discussion imageboards where vendors and customers were able to set up instances of face-to-face trading, typically with the assistance of instant messaging software such as Wickr or Telegram. The original, unmodified data set comprising 3,104,976 posts was collected and handed over to the ENNCODE consortium[1] by the site administration to be archived and shared for research purposes, as permitted by the site's Terms of Service. To promote the FAIR data principles, a reduced version of the corpus comprising 3,104,515 posts, referred to as the Finnish Dark Web Marketplace Corpus (FINDarC), has been deposited in the Language Bank of Finland, a language resource service coordinated by the national FIN-CLARIN consortium formed by Finnish universities and other research organizations. Researchers can contact the Language Bank and apply for permission to access the corpus under the CLARIN RES license offering a time-restricted personal license to re-use the data according to an approved research plan.[2]

While the dark web online market places, including Torilauta, emphasize user anonymity, the posts submitted to such sites can nevertheless contain personal information, such as unique usernames and personal names, enabling data subject re-identification. Therefore, as described in this paper, we have made our best effort to assess and identify the type and amount of personal information in the original unmodified data set, to assess and implement viable data anonymization/reduction approaches, to assess privacy and security measures implemented by the Language Bank of Finland, and to put in place a future corpus management plan coordinated by the Language Bank of Finland.

Those carrying out future research based on the corpus are encouraged to implement appropriate ethical proofreading measures (see e.g. (Harviainen et al., 2021)) in order to further mitigate any potential harm from access to the material, to both the researchers and the studied populations.

[1] Consortium website: https://research.tuni.fi/enncode/

[2] Permanent link to the corpus: http://urn.fi/urn:nbn:fi:lb-2022062221

2 Related work

Prior to the data publication presented here, Torilauta was utilized in the ENNCODE project as a data source in multiple linguistic and social science studies (Haasio et al., 2020; Hämäläinen et al., 2021; Hämäläinen & Ruokolainen, 2021; Harviainen et al., 2020; Karjalainen et al., 2021). In particular, Haasio et al. (2020) examined information needs of drug users using a sample of 9,300 posts.[3] Harviainen et al. (2020) studied cultural and socioeconomic aspects of drug traders using the same 9,300 post sample. Hämäläinen and Ruokolainen (2021) studied narcotic substance vocabulary based on a sample of 3,000 posts. Hämäläinen et al. (2021) studied a sample of 1,654 usernames extracted from posts submitted to the site. Karjalainen et al. (2021) examined the availability of illegal narcotics during the first wave of the COVID-19 pandemic using a sample of 535 posts.

It is notable that none of the previous studies attempted to share their data sets with the research community in a systematic manner. This practice negatively impacts the replication and verification of the published studies and potentially discourages further research on the topic. On the other hand, given the sensitive and potentially incriminating nature of the data, not releasing the data is an understandable approach since preparing and managing such a resource gives rise to multiple technical, ethical, and legal challenges. The purpose of this paper is to describe and discuss these challenges and how we approached them.

To the best of our knowledge, there exist relatively few published dark web corpora or text data sets. Three notable exceptions include the Dark Net Market archives (2013–2015) (Branwen et al., 2015), a collection covering 89 dark net markets and over 37 related forums (1.6TB uncompressed) scraped during 2013-2015, DUTA (Nabki et al., 2017), a set of 7,000 text samples formed by sampling the Tor network for two months, and CoDa (Jin et al., 2022), a set of 10,000 web documents tailored towards text-based dark web analysis. All three corpora comprise primarily English texts and are either publicly downloadable (Dark Net Market archives) or available to researchers upon request (DUTA, CoDa).

In a recent study, Leedham et al. (2021) discussed their work on archiving a hard-to-access WiSP corpus consisting of texts written by social work professionals describing their work practices. Due to the potentially sensitive nature of the texts, Leedham et al. (2021) created two versions of the corpus: one for the research project and an anonymized/reduced version for archiving. In a similar vein, our work presented here aims to provide an extensive discussion of the process of preparing a corpus of potentially sensitive texts for archiving and sharing.

3 Data set

Table 1: Data fields comprising a single post. The column titled *missing (%)* indicates the portion of all posts where the field value is not available.

	description	example	missing (%)
boardUri	board identifier	roi	0.0
creation	post creation datetime (UTC)	2020-01-14T17:51:24.714Z	0.0
deletion	post deletion datetime (UTC)	2020-01-27T16:49:03.663Z	2.8
threadId	thread identifier	27961	0.0
postId	post identifier	28069	29.8
name	poster name	example-name	54.1
subject	message subject	Example message subject	46.2
message	message text body	Example message text body	0.0

Each post in the corpus is represented as a data structure with 8 fields as shown in Table 1. The original data set received by the consortium included all posts submitted to Torilauta between 2019-09-11 and 2020-05-20 (1,863,639 posts in 251 days) and 2020-06-17 and 2020-10-31 (1,099,710 posts in 136 days). In addition to the posts collected during these active collection periods, the data contained "residue" posts submitted between 2017-11-02 and 2019-09-11 (141,627 posts in 678 days). Meanwhile,

[3] In their paper, Haasio et al. (2020) refer to *Torilauta* using its other commonly used name *Sipulitori*.

posts submitted between 2020-05-20 and 2020-06-17 were missing completely. Therefore, the original unmodified corpus consisted of 3,104,976 posts in total.

The data is grouped by boards. Of the 32 boards, the board with the highest activity measured by the total number of submitted posts and threads was the market board dedicated to narcotics transactions within the city of Helsinki (*/hki*). The total number of posts submitted to this board was 787,459 corresponding to 25.4% of all posts in the data. Meanwhile, in total 96.5% (2,997,624) of all posts were submitted to the 16 boards dedicated to transactions.

4 Data release

Text anonymization approaches proposed in the literature commonly utilize automatic named-entity recognition (NER) as a part of the processing pipelines to varying extents (Adams et al., 2019; Csányi et al., 2021; F. & Trabelsi, 2018; Francopoulo & Schaub, 2020; Garat & Wonsever, 2022; Glaser et al., 2021; Oksanen et al., 2019; Tamper et al., n.d.). Ideally, NER tools would also have been useful when processing FINDarC. However, examining the prediction quality on a manually annotated test section of the data set, suggested that the available tools suffered from a domain mismatch in addition to the inherent mismatch between personal data and named entity classes. This was not completely surprising since the text domain also caused problems for human annotators when creating the test section. Because the available tools tended to miss entities of interest (low recall) and be incorrect when detecting entities (low precision), we did not consider them efficient pre-processing tools for FINDarC in their current state.

Instead we decided to use full-text search, to find common personal identifiers with relatively rigid formats, such as social security numbers and phone numbers. We defined a target set of textual patterns (regular expressions) and searched for matches in message bodies. Specifically, we were interested in finding expressions matching 1) (Finnish) social security numbers, 2) (Finnish) phone numbers, 3) email addresses, 4) IBAN bank accounts, and 5) IP addresses, all of which have relatively rigid formats. We applied the search to all posts in the data and assigned the matches manually to personal data and non-personal data according to post context. We did not filter out noise from the data and instead applied the search to all 3,104,976 posts in the original corpus.

As shown in Table 2, the most and least frequent matched types were email addresses and bank account numbers with 1,840 and 12 regular expression matches, respectively. Due to the sufficiently low number of original matches, we were able to perform manual verification of all the cases.

Table 2: Matched regular expression frequencies. The columns titled *matches* and *verified* denote the number of found regular expression matches and the number of manually verified cases, respectively, The columns titled *posts* and *threads* denote the number of distinct posts and threads where the verified cases occurred.

	matches	verified	posts
phone	875	858	699
hetu	91	73	65
email	1,840	1,837	1,707
iban	12	12	12
ip_address	121	16	14
total	2,939	2,796	2,261

The phone numbers and email addresses occurred in two contexts. First, similarly to the instant messaging usernames , 491 out of 858 and 1,622 out of 1,837 of the phone numbers and email addresses, respectively, were posted as contact information by the individuals themselves. The remaining cases were posted as a means of targeting people. In such cases, personal details (e.g., name, relationship information, area of residence) were shared in connection with one or more usernames, in order to paint the person as a potential target for violence. Bank account numbers occurred similarly in two contexts. Out of the 16 IP addresses, 10 cases were included as a means of targeting, while the remaining 6 were provided as a type of contact information. Finally, all 73 and 12 found cases of social security numbers and bank account numbers were posted with a purpose of targeting. Thus, we identified and removed in total

667 cases of targeting in 295 posts using this method. Finally, we created a second regular expression list using words and prefixes related to the personal information contained in the identified 295 targeting posts. This list consisted of 77 keywords and parts of person names and addresses.[4] After performing a second search with these patterns and a subsequent manual inspection, we identified and removed an additional set of 166 posts submitted as a means of targeting.

5 Protective measures

Conventionally, the most direct approach to protect data subjects from re-identification has been to anonymize the data by removing/obscuring the parts containing personal information (Ohm, 2009). However, it appears evident that, if implemented successfully, this type of processing would have a profound impact on the usefulness of FINDarC for research purposes. For example, subsequent to removing usernames from their post contexts or from the data altogether, one would not be able to replicate the study of Hämäläinen et al. (2021) who examined how sellers and buyers of illegal drugs represent themselves in their usernames. In turn, subsequent to removing location and/or timestamp data, one would no longer be able to replicate the study of Karjalainen et al. (2021) who studied the availability of drugs specifically in the city of Tampere during the COVID-19 epidemic in the spring of 2020. From a utility point of view, therefore, it could be argued that reducing personal information from the buy/sell post threads would quickly degrade, or destroy, the usefulness of the corpus as a data source for research. This problem is generally referred to as the privacy-utility trade-off within the data privacy literature (Alvim et al., 2011; Li & Li, 2009). If anonymisation is not an option, there are, however, other means to protect the data that need to be justified with the help of a data impact assessment.

Due to the problematic privacy-utility trade-off, we posit here that reducing the FINDarC extensively would not be appropriate even if sufficient resources could be allocated for domain-specific tool development and manual labour. Furthermore, we note that Torilauta and other drug trading sites have also been under observation by other parties, including both criminals and law enforcement agencies. Therefore, it is our assessment that leaving the sell/buy posts, which form the majority of the FINDarC, largely intact poses few additional risks to the studied populations as they had entered their data for public use. However, in addition to the sell/buy posts, the data also contains posts with the intention of doxxing/targeting individuals. Here, our position is that removing these submissions is warranted from an ethical point of view while not decreasing the value of the corpus as a data source significantly. This is because these posts are not directly related to the main functionality of the site as an online marketplace. Accordingly, we removed from the corpus all 461 posts containing identified doxxing/targeting information The reduced corpus, therefore, comprises 3,104,515 posts.

Finally, as per the Terms of Service of Torilauta, the site users gave consent to data collection for academic use by using the site. Consequently, site users could opt out of the data collection by not submitting new posts and/or contacting the site administration about previously submitted posts. However, it could be argued that by removing a previously submitted post, a user has withdrawn the permission to use the data. Unfortunately, the original data set received from the site administration did not include information about the reasons behind post deletions. Therefore, we were not able to exclude any posts from the corpus based on the deletion status.

Due to the limited applicability of data reduction as a means of protecting data subjects from re-identification, we instead need to restrict the access to the corpus. Since the FINDarC resource in its current form contains personal data, both copyright and personal data legislation apply and the corpus cannot be published with open access. Instead, FINDarC has protected access under the CLARIN RES licence which means that permission to download and use the corpus is only granted to researchers based on written applications reviewed by the data controller (principal investigator of the ENNCODE consortium) including a data protection impact assessment. The purpose of this limitation is to ensure that the material is accessed only by verified researchers for legitimate research purposes. It also lessens sharing-related risks to both the researchers and the subjects of study, as mandated by the consortium's data management policy.

[4] We do not present the list here due to obvious privacy issues.

Restricting access to the FinDARC corpus as described in this paper is in line with the current literature on personal data sharing (Elliot et al., 2020; Elliot et al., 2018; Ohm, 2009; Rubinstein & Hartzog, 2016; Stalla-Bourdillon & Knight, 2016) which adheres to the FAIR principles, while acknowledging the limitations of data anonymization/reduction and encouraging the use of user group limitations.

6 Conclusions

We have discussed the archiving procedure of FINDarC, a Finnish dark web marketplace corpus, in the Language Bank of Finland. It was unlikely that the corpus could be fully anonymized to be shared publicly without also compromising its value for research, so instead other protective measures were taken to make it possible to share the data. The discussion included an overview of the data, assessment of the risk and impact of data subject re-identification, assessment and implementation of viable data reduction approaches using manual and automatic text processing, assessment of privacy and security measures implemented by the Language Bank of Finland, and a future corpus management plan implemented and coordinated by the Language Bank of Finland. As a result of the presented work, a reduced version of the corpus has been archived in the Language Bank of Finland. Researchers can apply for access to the corpus under the CLARIN RES licence.

Acknowledgments

We acknowledge the funding for the Language Bank and FIN-CLARIN by the Academy of Finland.

References

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., & Wechsler, R. (2019). AnonyMate: A toolkit for anonymizing unstructured chat data. Proceedings of the Workshop on NLP and Pseudonymisation, 1–7.

Alvim, M. S., Andrés, M. E., Chatzikokolakis, K., Degano, P., & Palamidessi, C. (2011). Differential privacy: On the trade-off between utility and information leakage. In International workshop on formal aspects in security and trust (pp. 39–54).

Branwen, G., Christin, N., Décary-Hétu, D., Andersen, R. M., StExo, E. P., Anonymous, D. L., Sohhlz, D. K., Cakic, V., Buskirk, V., Whom, M. M., & Goode, S. (2015). *July* [Dark net market archives, 2011-2015.]. %5Curl%7BAccessed:%202022-06-28%20https://www.gwern.net/DNM-archives%7D

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., & Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. Symmetry 13*(8): 1490*.

Elliot, M., Mackey, E., & O'Hara, K. (2020). In *The anonymisation decision-making framework 2nd Edition: European practitioners' guide*.

Elliot, M., O'hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law & Security Review*, *34(2)*, 204–221.

F., D. C., & Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. *In European Conference on Advances in Databases and Information Systems*, 118–126.

Francopoulo, G., & Schaub, L. P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. *workshop on Legal and Ethical Issues (Legal2020)*, 9–14.

Garat, D., & Wonsever, D. (2022). Jan. Automatic Curation of Court Documents: Anonymizing Personal Data. In *Information 2022, vol. 13* (pp. 1–27). https://doi.org/\doi{10.3390/INFO13010027}

Glaser, I., Schamberger, T., & Matthes, F. (2021). Anonymization of German legal court rulings. In *Proceedings of the 18th international conference on artificial intelligence and law, icail 2021* (pp. 205–209). https://doi.org/\doi{10.1145/3462757.3466087}

Haasio, A., Harviainen, J. T., & Savolainen, R. (2020). Information needs of drug users on a local dark Web marketplace. *Information Processing and Management*, *57*(2), 1016. https://doi.org/\doi{10.1016/j.ipm.2019.102080}

Hämäläinen, L., Haasio, A., & Harviainen, J. T. (2021). Usernames on a Finnish Online Marketplace for Illegal Drugs. *Names - A Journal of Onomastics*. https://doi.org/\doi{10.5195/NAMES.2021.2234}

Hämäläinen, L., & Ruokolainen, T. (2021). Kukkaa, amfea, subua ja essoja: Huumausaineiden slanginimitykset Tor-verkon suomalaisella kauppapaikalla. *Sananjalka*, *63*, 130–153. https://doi.org/\doi{10.30673/sja.106615}

Harviainen, J. T., Haasio, A., & Hämäläinen, L. (2020). Drug traders on a local dark web marketplace. *ACM International Conference Proceeding Series*, 20–26. https://doi.org/\doi{10.1145/3377290.3377293}

Harviainen, J. T., Haasio, A., Ruokolainen, T., Hassan, L., Siuda, P., & Hamari, J. (2021). Information protection in dark web drug markets research. *Hawaii International Conference on System Sciences*.

Jin, Y., Jang, E., Lee, Y., Shin, S., & Chung, J. W. (2022). *Shedding new light on the language of the dark web* [*arXiv preprint (To appear in NAACL 2022)*].

Karjalainen, K., Nyrhinen, R., Gunnar, T., Ylöstalo, T., & Ståhl, T. (2021). Huumeiden saatavuus, käyttö ja huumausainerikollisuus Tampereella koronakeväänä 2020. *Yhteiskuntapolitiikka*, *86*(2), 80–90.

Leedham, M., Lillis, T., & Twiner, A. (2021). Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP Corpus. *Applied Corpus Linguistics*, *1*(3). https://doi.org/10.1016/j.acorp.2021.100011

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 517–526.

Nabki, A., M. W., E. F., Alegre, E., & Paz, I. D. (2017). Classifying illegal activities on tor network based on web textual contents. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 35–43.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701.

Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., & Hyvönen, E. (2019). AnoPpi: A pseudonymization service for Finnish court documents. In *Jurix 2019* (pp. 251–254). IOS Press.

Rubinstein, I. S., & Hartzog, W. (2016). Anonymization and risk. *Wash. L. Rev*, *91*, 703.

Stalla-Bourdillon, S., & Knight, A. (2016). Anonymous data v. personal data-false debate: an EU perspective on anonymization, pseudonymization and personal data. *Wis. Int'l LJ*, *34*, 284.

Tamper, M., Oksanen, A., Tuominen, J., Hyvönen, E., & et. al., A. H. (n.d.). Anonymization Service for Finnish Case Law: Opening Data without Sacrificing Data Protection and Privacy of Citizens. *International Conference on Law via the Internet, LVI*.

The making of the CLARIN Resource Family for Oral History: Lessons Learned from 'Voices of Ravensbrück'

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

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

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

Christoph Draxler
Ludwig Maximilian University, Munich, Germany
christoph.draxler@lmu.de

Abstract

The outcome of two CLARIN funded projects in 2021 and 2022 for curating a set of oral history collections was the introduction of a new type of CLARIN Resource Family (CRF) in 2023: *Oral History*. In this paper we elaborate on the added value of this category for the CLARIN research community, we describe how we created the *Oral History* CRF *Voices of Ravensbrück*, and we draw conclusions on the feasibility of standardising interviews coming from different institutional contexts to create a CLARIN Resource Family.

1 Introduction

This paper reflects on the creation of a new type of CLARIN Resource Family, namely *Oral History*. Data originating from a historical discipline, oral history, was brought together with the aim of offering it to an audience of linguists and social scientists. Although several linguists have analyzed oral history interviews by applying linguistic tools (e.g., Salah et al. 2021, Gerstenberg, Pagenstecher 2022, Mlynář 2022), the use of interviews in the realm of language studies is relatively new. Interviews that have been collected for other purposes and are reused can be considered as *legacy data*, "recordings made at any time in the past […] as opposed to new recordings made in the field or the laboratory in the course of a new study" (Bounds, et al., 2011: 46). Progress in digital and web- technology has facilitated the transition from analogue interviews on cassettes or tape in the archive to digital records enriched with metadata that are retrievable on the web. Speech retrieval technology, interface design, automatic metadata extraction and translation have smoothed the path for potential international collaboration and comparison of oral history data. Especially historical phenomena that involve the large scale (forced) movement of people from one geographical area to another, such as migration, war and diaspora, offer ample possibilities for transnational research. Our goal was to lay the basis for a transdisciplinary and transnational interpretation of experiences in camp Ravensbrück, the only female only concentration camp that was run in Nazi-Germany during World War II.

The starting point for the project was the discovery of a collection of compact cassette tapes with interviews on concentration camp Ravensbrück by linguist Silvia Calamai. These were held by Italian writer Anna Maria Bruzzone with five Italian Ravensbrück survivors back in 1977, to collect material for her book *Le donne di Ravensbrück* (Beccaria Rolfi, Bruzzone 1978, 2020). Bruzzone's niece donated the analogue collection to Siena University, where it was digitised and added to the archive of the University Library. Aside from receiving the consent of the next of kin, publishing the data was possible

because of an Italian law stating that access to such data must be provided in the legitimate interest of research forty years after the data has been created (Legislative Decree 22/01/2004, art. 122).

Considering the unique multilingual context of the camp with female prisoners coming from all countries occupied by Nazi-Germany, our intention was to collect a multilingual set of interviews that could yield interesting research scenarios for memory studies, trauma studies, gender studies, archival studies, linguistics and oral history.

With the support of CLARIN, the Italian collection of five interviews was curated in order to meet the criteria for Open Access and to be accessible to an international audience. In addition, 36 interviews with survivors from Ravensbrück conducted in English, German and Dutch that are retrievable through the web, were identified in several online archives, project web sites, and repositories. Together with the Italian interviews they form the CLARIN Resource Family *Voices of Ravensbrück*[1]. For the complete set of interviews, the corresponding metadata was collected and converted into the CMDI (component metadata infrastructure.

2 Experiences with Bruzzone's interviews: from transcription to dissemination

The digitized Bruzzone recordings (approx. 20:15h of recorded speech in Italian) were transcribed as a verbatim orthographic transcript, adhering to the requirements of linguistic research. This includes documenting elements such as non-linear exchange (e.g., speech overlaps or abnormal turn-takings), hesitations and reformulations and other salient acoustic signals (e.g., interjections, laugh). The transcripts have also been time-aligned at the phrase level. Furthermore, with an international audience in mind, detailed summaries in English were created of each interview. The transcription was performed by a native speaker of Italian at Siena University using the transcription editing tool Octra in local mode.

This tool is part of the Open Source transcription portal the *T-chain*[2] that was developed with the support of CLARIN, but it can also be used separately. The portal features access to third party ASR providers (e.g., Fraunhofer, Radboud University, Google, IBM) via the web services of the Bavarian Archive for Speech Signals in Munich[3]. The expectation was that by using the portal and the Italian language model, the transcription process could be speeded up. Yet, it turned out that correcting the ASR results costed more effort than creating the transcripts from scratch.

We concluded that the performance of ASR depends on elements that cannot be fully controlled with *legacy data*: 1. the quality of the audio signal, 2. a clear separation of speaker turns, and 3. the availability of a language model for the given domain of discourse (see Draxler 2022 for a discussion). In addition, requirements for linguistic annotation such as speaker turns, hesitations, self-repairs, backchannel feedback, or salient phenomena such as laughing, have a severe impact on transcription speed. In sum, the time-consuming tasks of transcribing and making linguistic annotations, still needed to be performed by a human expert.

As could be expected from such a unique finding – a set of original interviews with Italian survivors of camp Ravensbrück – the end results of the project were well received, especially in Italy. A CLARIN café was organised to present the main results to the research community and to all data providers[4], during which the President of the *Comité International de Ravensbrück*, Ambra Laurenzi, presented the activities of this organisation. Anticipating the expansion of the project in the near future to Eastern Europe, Alessandra Carbone and Yuliia Chernyshova from Siena University presented their survey on Ukrainian, Russian, and Polish oral history data on Ravensbrück (in their survey also the Ukrainian Canadian Research & Documentation Center[5] located in Canada was involved). The interest from the side of the media was extensive: a lengthy article appeared in the Italian daily newspaper *Corriere della Sera – La lettura* (21.01.2023) and Silvia Calamai was interviewed by a regional TV program and a monthly magazine (see overview press release by Siena University on the occasion of International Holocaust Remembrance Day 2023)[6]. Stefania Scagliola and Silvia Calamai were also invited to an

[1] www.clarin.eu/resource-families/oral-history-corpora
[2] clarin.phonetik.uni-muenchen.de/apps/TranscriptionPortal/
[3] clarin.phonetik.uni-muenchen.de/BASWebServices/
[4] www.clarin.eu/event/2022/clarin-cafe-voices-ravensbruck-web-multilingual-challenge
[5] www.ucrdc.org/
[6] www.unisi.it/unisilife/notizie/voci-da-raversbruck-le-risorse-line

international workshop held at Yad Vashem, on new ways of presenting the Holocaust with digital tools (Yad Vashem, Jerusalem, 22-24 November 2022). Finally, the European Holocaust Research Infrastructure (EHRI) and CLARIN decided to combine their expertise to explore Oral Testimonies of the Holocaust in a Hackathon that was held at King's College London in May 2023[7].

3 Experiences with identifying and sharing oral history interviews about Ravensbrück in different languages

Putting together a dataset of interviews on camp Ravensbrück, conducted in different languages and created in different institutional contexts proved to be quite a challenge. Comparing the role of language in the camp in the Italian interviews with how this theme is addressed in the other interviews was hard to realize. Partly because it was not prioritized in the questions raised during the interviews, partly because of the structural lack of metadata and transcripts in the online archives. An interesting finding though, when comparing the languages spoken in the camp with those spoken during the interviews, was the role of migration. The Italian dataset was exceptionally consistent: Italian women are deported to a German camp, isolated as a minority group, return to Italy after they are liberated, and decades later are all interviewed in Italian by an Italian woman of their age for an Italian publication. This consistency was absent in the 36 interviews that we could identify in English, German and Dutch, within large scale collections about the Holocaust in the United States (United States Holocaust Memorial Museum, Fortunoff Collection, Shoah Visual History Collection). It is striking how interviews with Ravensbrück survivors are held in English with women of Polish or Russian origin who migrated to the US, or in Hebrew with women who migrated to Israel. Because of the Anglo-saxon dominance in online oral history resources, these large scale collections can be regarded as the most influential with regard to camp experiences. The two major projects that specifically deal with women and Ravensbrück are to be found in Germany (Loretta Walz's archive) and Austria (Österreichische Mediathek). The first is multilingual, but all metadata is translated into German. The second focuses on Austrian survivors of Ravensbrück, and offers metadata in German. A paper by Calamai & Scagliola on this topic is in preparation.

Our plan to curate the 36 recordings and send them back to their owners enriched with missing metadata, or/and with alignment of text and sound, with the help of the Open Source transcription tool T-Chain, was too ambitious, due to copyright restrictions and institutional hurdles. This is also related to a different research practice with regard to the principle of 'access' to a sound recording. In linguistics it speaks for itself that annotating or analysing a sound or video file, requires getting hold of it on one's own computer, or in a virtual research space. For oral historians online 'access' means enabling listening or viewing the content and presenting it with an interface that allows exploring the archive. A download option is often limited to a transcript. Enabling the download of a recording of the narrator's personal story is seen as a step too far. Linguists usually use spontaneous speech, recorded with no aim to document personal stories about someone's private life, while oral history interviews are bound to a living person or to his or her next of kin (Calamai in press). Moreover, to perform source criticism and offer a proper interpretation of the interview, a scholar needs to know where, when and how the interview was conducted. For linguists the essence of the data is the speech itself, not the background and personal history of the narrator.

In addition some curators of oral history collections are concerned that the first generations of narrators ('60, '70,'80, '90) could not give consent to be directly accessible on the web. Another hurdle is financial. Collections that are not publicly funded need a business model to sustain their work, and cannot be shared freely. The main hurdle however, was backlog in curation. As long as this persists, it will be difficult to connect data from different institutional contexts.

In view of these limitations we followed the following procedure. After identifying interviews and getting permission from the copyright owners, we requested an export of the metadata and included the URL of the online archive where one can request an account. These metadata were then copied into the CMDI metadata scheme. (see the next section). Thus we enhanced the findability of and the information about these interviews for the CLARIN community.

[7] www.clarin.ac.uk/article/using-holocaust-testimonies-research-data

4 Experiences with metadata integration into the VLO

The first lesson we learned with regard to converting the existing metadata fields of the Bruzzone collection into the CMDI scheme that is current in CLARIN, is that legacy data requires more background information in the metadata. This is why an extended and slightly modified version of an earlier profile[8] was created *OralHistoryInterviewCRF*[9], in the CMDI component registry. Two components were added: context of creation and context of digitisation. Moreover, the metadata set is more detailed on interview content and method than in standard spoken corpora. For all the interviews of the CRF *Voices of Ravensbrück*, the records of this metadata profile were populated using the COMEDI metadata editor[10] (Lyse, et al., 2015). These records were then downloaded from COMEDI, and copied with a selflink to the CLST's OAI-PMH endpoint, from which they can be harvested in CLARIN's Virtual Language Observatory[11] (VLO). The full process is demonstrated in a video[12]. The Bruzzone interviews and their transcripts are stored and shared with restricted access through The Language Archive[13].

5 Conclusion

Going through the process of translating different oral history metadata schemes to the CMDI metadata scheme is demanding for scholars outside the realm of linguistics. To support the 'translation' of the jargon and structure of a CMDI metadata scheme we created a document with an extensive description of each of the components of our CMDI OH profile[14]. This can be found on the landing page of the CRF through this link. With regard to the structure of the profile we recommend developing a workflow at the very start of the project, showing the required metadata and the hierarchy of its components, and how the metadata can be integrated into the VLO. Departing from a standard spreadsheet with required metadata elements leaves too much room for differences in interpretation. Visual clarity also plays a role in the design of the COMEDI metadata-editor, that is designed in such a way that fields with missing values do not appear in the final record. This makes it difficult to compare records and assess which fields are missing.

With regard to our objective of creating a transnational and transdisciplinary dataset with oral history interviews on camp Ravensbrück, we gained much knowledge on the diversity of archival structures and procedures for access among oral history collections that are retrievable on the web. This diversity is related to copyrights, different interpretations on the principle of 'access' and backlog in the creation of metadata and transcripts. The latter could not be compensated by processing the audio with the T-chain, our transcription portal, either because we could not get hold of the audio-files, or because the quality of the legacy data was insufficient for a good performance of the ASR.

To reach the level of interoperability that technology promises us between oral history records coming from different institutional contexts, there is still a long and winding road to go. The Italian records, however, offered us proof that the gap between oral history and linguists can be bridged. The study of the relation between the politics of oral history, gender, memory and language deserves further research.

References

Bounds, P., Palosaari, N., William, A. Kretzschmar, Jr. (2011). Issues in using legacy data. In Di Paolo, M. Yaeger-Dror M. (eds.), *Sociophonetics: A student's guide*. Routledge, 46-57.

Calamai, S. in press. Oral history and sociophonetics, in Ch. Strelluf (ed.), *Routledge Handbook of sociophonetics*, Routledge.

Calamai, S., Scagliola, S., Ardolino, F., Draxler, Chr., Van Hessen, A., Van den Heuvel, H. (2022). Ravensbrück Interviews: How to Curate Legacy Data to Make it CLARIN Compliant. Selected Papers from the CLARIN

[8] catalog.clarin.eu/ds/ComponentRegistry/#/?itemId=clarin.eu%3Acr1%3Ap_1369752611609&registrySpace=public
[9] catalog.clarin.eu/ds/ComponentRegistry#/?itemId=clarin.eu%3Acr1%3Ap_1653377925732&registrySpace=public
[10] clarino.uib.no/comedi/page
[11] www.clarin.eu/content/virtual-language-observatory-vlo
[12] www.youtube.com/watch?v=1ePB5H31GKs&t=1825s&ab_channel=CLARINERIC
[13] archive.mpi.nl/tla/islandora/object/tla:1839_76375cde_a68e_4c87_8539_513c3a63e308
[14] docs.google.com/document/d/115UH0s0uP3t1P7zuBoTOmXH8XUAQmyNmjrIzzkUfFv4/

Annual Conference 2021. Series: *Linköping Electronic Conference Proceedings 189*. ISBN: 978-91-7929-444-1. ISSN: 1650-3686 (print), 1650-3740 (online). DOI: https://doi.org/10.3384/9789179294441

Draxler, Chr. (2022). Automatic Transcription of Spoken Language Using Publicly Available Web Services. In: *Fare linguistica applicata con le digital humanities*, Studi AItLa, 14, 27-49. http://www.aitla.it/images/pdf/StudiAItLA14/ebookAItLA_14.pdf

Gerstenberg, A., Pagenstecher, C. (2022). "Mi ricordo", "je me souviens": ich erinnere mich. Sammlungsübergreifende Interviewanalysen, *Oral History und Korpuslinguistik. apropos: Perspektiven auf die Romania*, 9, 213-239. DOI: https://doi.org/10.15460/apropos.9.1902

Lyse, G.I, Meurer, P. & De Smedt, K. (2015). COMEDI: A component metadata editor. *Proceedings. Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*. Linköping Electronic Conference Proceedings 116:8, 82-98. https://ep.liu.se/ecp/116/008/ecp15116008.pdf

Mlynář, J. (2022), How is Oral History Possible? On Linguistically Universal and Topically Specific Knowledge. *The oral history review*, 49, 1, 116-132 https://doi.org/10.1080/00940798.2022.2050412

Salah, A.A., Salah, A.A., Kaya, H., Doyran, M. Kavcar, E. (2021), The sound of silence: Breathing analysis for finding traces of trauma and depression in oral history archives. *Digital Scholarship in the Humanities*, 36(2) https://doi.org/10.1093/llc/fqaa056

The LiRI Corpus Platform

Jonathan Schaber, Johannes Graën, Daniel McDonald, Igor Mustač,
Nikolina Rajović, Gerold Schneider and Noah Bubenhofer
Linguistic Research Infrastructure
University of Zurich, Switzerland
first_name.last_name@linguistik.uzh.ch

Abstract

We present the LiRI Corpus Platform (LCP), a software system and infrastructure for querying a vast array of corpora of different kinds. It heavily relies on the PostgreSQL relational database management system, employing state-of-the-art data representation and indexing techniques, which lead to significant performance gains when querying, even for structurally complex queries involving nested logical operations and quantifiers. In this work, we provide details on a number of said techniques and describe our approach to query representation and transformation.

1 Introduction

Corpora are an important resource for empirical linguistic research, as well as text-based social studies such as digital humanities or media research (Meurers, 2005). Already, there exist numerous tools and platforms for manipulating different corpus types (mono- vs. multilingual, text vs. multimodal, diachronic, etc.), offering various means of access to corpora (web interface, command line interface, etc.), and providing different functionalities (editing, querying, annotating, etc.).[1]

Most tools and applications developed thus far are designed with a specific corpus or corpus type as "target" and therefore lack a certain ability to generalize. Some tools provide a general interface for diverse corpora, handling a vast array of differently structured corpora and occasionally providing specific functionality needed for them (see section 2 for a (small) overview of existing tools and applications).

Central to all corpus tools is the ability to query one or more datasets. We propose a distinction between two main querying strategies:

1. **Corpus exploration**, characterized by being an iterative, drilling-down process, where users run and refine their queries based on results, and by interfaces that prioritize speed (Hearst, 1999).

2. **Corpus analysis**, maximizing recall of results in a corpus, prioritizing accuracy and specificity, with results returned in a format suited for subsequent statistical analyses.

The latter strategy is of particular interest to approaches that require large amounts of data (Big Data), such as the training of large language models.

While those two strategies are more complementary than antagonistic, to our knowledge most tools tend to prioritize the latter strategy over the former (Desagulier, 2019). As we show in Section 3, we design our application to accommodate both approaches, supporting the construction of queries in a bottom-up fashion. We see the incorporation of both strategies as a timely improvement over existing tools that presume their users to arrive with full-fledged, complex queries at the tool.

In this contribution, we present the LiRI Corpus Platform (LCP) a new tool that couples complex analysis of diverse corpora and interoperability with existing CLARIN resources. The software facilitates access to and re-use of research data, offering a flexible architecture, whereby a single backend can be connected to multiple frontends tailored to the requirements of specific research agendas. We will mostly focus on the data model and its implementation in an RDBMS, while only briefly explaining the different

[1] https://corpus-analysis.com/, a page collecting tools for corpus linguistics, lists 266 entries at the last date of access (September 24, 2023).

interfaces that we plan to implement. While interfaces are to some extend interchangeable, the dynamic data model and its implementation form a core feature and contribution of our tool.

2 Related Work

Myriad systems for storing and querying large corpora have been developed in past decades. Clematide (2015) gives an overview of corpus linguistic query language types, distinguishing between four families. Historically prominent were (I) sequence-based designs, such as CQP (Christ, 1994; Evert & Hardie, 2011) and other dialects of regular expressions, (II) structure-based designs, such as TGrep2 (Rohde, 2005) for querying syntactic trees. Many of them have organically developed within parameters set by technological restrictions of the time. For example, CQP does not allow syntactic queries because its sequence-based conceptualization of text is ill-suited to express non-sequential structural information. Extensions, notably CQL (Jakubíček et al., 2010) known from SketchEngine implement limited syntactic queries by means of its function `within`. More recently (III), the class of path-based based languages which use the XPath query language have been implemented. Finally, (IV) the class of logic-based languages such as TigerSearch and ANNIS, which offers outstanding expressiveness, coupled with considerably longer retrieval times. For example, for performance reasons, the most recent major version of ANNIS (Krause & Zeldes, 2016) has abandoned the relational database PostgreSQL, and developed a custom implementation based on graphs[2]. In contrast, our proposed approach retains the expressiveness of logic-based languages while leveraging advanced data-representation and indexing techniques in order to offer faster retrieval times.

Several methods for speed-ups have been proposed. For instance, while older approaches rely on MapReduce techniques (Schneider, 2013), modern database management systems as PostgreSQL provide internal mechanisms for parallel computing. Other methods are the use of sophisticated indexing and retrieval techniques, as in the the proposals of (Ghodke & Bird, 2012).

For reasons of space, our overview provides only a brief sketch of past and present resources. For example, we omit detailed description of several projects fine-tuned for particular corpora and particular tasks, for example Dependency Bank (Lehmann & Schneider, 2012), which allows fast syntactic queries on the British National Corpus (BNC).

While some published standards and guidelines have attempted to define a single digital format able to encompass all possible linguistic data and annotations (Gries & Berez, 2017; Ide & Romary, 2004, 2006), to our knowledge no actual working implementation of such proposals has been presented to date. While we do not suggest that our solution covers all possible needs, we offer an attractive solution that facilitates analysis of a very broad range of corpus types.

3 Central Methods and Technology

We offer a flexible solution that we have tested on several types of corpora, ranging from digital editions (where access back to the original page picture is essential), multimodal corpora (where movies and temporal annotation are required), and very large corpora (where short retrieval times are a priority, even for metadata-rich datasets).

The corpora that we are making available through the system differ in various aspects such as size, annotation layers and complexity. While structurally complex corpora pose a challenge for the construction and processing of a dynamic query language, very large corpora, on the other hand, demand the efficient retrieval of previously unseen queries. As to structurally complex corpora, we allow queries on parallel corpora by representing alignment as structural element, similar to dependency relation.[3] We employ a modern database management system (PostgreSQL) to deal with the latter challenge. Unlike the developers of ANNIS (Krause & Zeldes, 2016), who replaced PostgreSQL as a storage and query backend

[2] https://raw.githubusercontent.com/korpling/ANNIS/main/CHANGELOG.md (September 24, 2023).

[3] There are three main differences between these two relations: 1) Alignments represent correspondence and are thus undirected, alignment links are typically unlabelled, and 3) the data type of alignments are raster (nested) sets than trees (Graën, 2018).

with their own solution[4]

Our description of methods focuses on the central task of querying. Our methods have been designed to scale to very large corpora with complex annotation schemes. To this end, we have designed a query definition representation in JSON format. We can translate queries from query languages like CQL to that format, but the other way round, there are queries which we can represent in that format, but we cannot represent in CQL (e.g. queries using syntactic relations or alignment structures).

Since the architecture of LCP is quite complex and is intended to serve as a platform for making linguistic data available e.g. through the CLARIN-network, it is implemented as a service that can be accessed over the internet; either through several dedicated web applications or directly via an API. Both access method are secured by requiring a login through an identity provider.

In this section, we limit ourselves to two central techniques that we have successfully employed in our infrastructure. However, there are a multitude of other improvements we cannot explain in detail here due to limited space.

3.1 Logarithmic Partitioning

LCP places no inherent restrictions on corpus size.[5] Querying a multi-billion word corpus can potentially take prohibitive amounts of time, prohibiting a seamless iterative querying approach, despite this being a typical workflow in corpus-driven linguistic research (Rayson et al., 2017). To quickly retrieve initial results, and thus enable the researcher to refine their query and estimate the distribution of the phenomenon on the whole dataset, we propose a structure which allows incrementally querying randomized subsets of the corpus data.

To this end, we randomly generate Universally Unique Identifiers (UUIDs) (Leach et al., 2005) as primary keys for the linguistically salient units in each corpus – typically, sentences or utterances. We then subdivide the total space of potential IDs into partitions of decreasing size, always splitting one part into two halves. Since Generation 4 UUIDs are generated by a truly random algorithm, we can expect each partition to comprise a known share of the total number of units, that is, a half, a fourth, and eighth etc.

In the case of a one billion word corpus with an estimated average of 10 words per sentence and a needed one million words for the smallest partition, we would thus create $100\,000 = 100\,000\,000/2^x$, that is $x = \log_2 1\,000$, that is $x \approx 10$, i.e. 10 partitions.

This logarithmic partitioning allows us to run a query on the smallest partition and extrapolating from the first result set retrieved both the expected amount of results on the whole corpus and the best next partition to query in order to satisfy a request for a particular number of results.

For the investigation of frequent phenomena, a small random sample is often enough. If more data is needed, our approach seamlessly scales from a quick 'pilot study' on a subset of the corpus to a complete analysis of the entire dataset. Such pilot studies also allow researchers to test and debug their queries quickly, and to assess where interesting differences by the available metadata can be found (period, variety, genre, gender, etc.) in an exploratory fashion.

3.2 Indexable Vector Representations

Corpora added into our database are passed through a pipeline that computes a vector representation of each sentence, preserving the positional information of tokens; in PostgreSQL this data structure is implemented under the name of "tsvector" (Bartunov & Sigaev, 2001). This is related to the classic information retrieval task of phrase searches (Manning, 2009). In most text corpora, tokens show various layers of annotation (e.g. lemmas, part-of-speech tags, morphological features etc.). With the vector representation implemented by that data structure, we can define multiple pieces of information per position (in a sentence). In order to take them apart when querying, we need to distinguish word forms

[4] "Instead of using the relational database PostgreSQL, a custom AQL implementation based on graphs called graphANNIS is used." (https://raw.githubusercontent.com/korpling/ANNIS/main/CHANGELOG.md (September 24, 2023)).

[5] This is one of the advantages of LCP over other systems like e.g. the Corpus Workbench, whose architecture restricts it to a maximum corpus size: "[...] the internal format of the indexed and compressed corpora imposes [a] 2.1 billion token limit" (Evert & Hardie, 2011, p. 13).

from lemmas, tagsets and so on. To this end, we prepend each string with one character per layer to tell them apart.

The advantage of such a vector structure is that it can be efficiently indexed[6] and thus allows performant retrieval of sentences that match specific criteria. Typical search patterns like sequences of tokens with additional constraints on annotations, like a CQP-style `[pos="DET"] [pos="ADJ"] [lemma="linguist"]` can be converted to a vector query that makes use of this index and quickly finds matching sentences. This allows us to often drastically reduce the number of sentences that need to be further filtered.

4 Discussion

We describe a nascent system that successfully adapts innovative structural and indexing solutions, using a modern database in order to handle complex queries of large linguistic corpora and performant retrieval of query results. Its indexing techniques lead to significant performance gains when querying, even for structurally complex queries. Our strategy of logarithmic partitioning means that first trends from the first partitions are available very fast, which allows prototyping on corpora of nearly unlimited size. The data model we employ in the database is designed to be as flexible as possible, allowing the representation of various corpora having very different, possibly deeply nested hierarchical structures (books, chapters, verses, parliamentary discussions, etc.).

At the moment, we provide an interface for the automatic upload of simple-structured ConLL-U adhering corpora; however, this will be expanded in the future and allow the uploading of more complex and multi-modal corpora.

References

Bartunov, O., & Sigaev, T. (2001). *Full-Text Search in PostgreSQL* (tech. rep.).

Christ, O. (1994). A Modular and Flexible Architecture for an IntegratedCorpus Query System. *arXiv preprint cmp-lg/9408005*.

Clematide, S. (2015). Reflections and a Proposal for a Query and Reporting Language for Richly Annotated Multiparallel Corpora. In G. Grigonyte, S. Clematide, A. Utka, & M. Volk (Eds.), *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015* (pp. 6–16). Linköping University Electronic Press.

Desagulier, G. (2019). Can word vectors help corpus linguists? *Studia Neophilologica*. https://shs.hal.science/halshs-01657591v2

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium.

Ghodke, S., & Bird, S. (2012). Fangorn: A system for querying very large treebanks. *Proceedings of COLING 2012: Demonstration Papers*, 175–182.

Graën, J. (2018). *Exploiting alignment in multiparallel corpora for applications in linguistics and language learning* (Doctoral dissertation). University of Zurich. https://doi.org/10.5167/uzh-153213

Gries, S. T., & Berez, A. L. (2017). Linguistic Annotation in/for Corpus Linguistics. *Handbook of linguistic annotation*, 379–409.

Hearst, M. A. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual meeting of the Association for Computational Linguistics*, 3–10.

Ide, N., & Romary, L. (2004). International standard for a linguistic annotation framework. *Natural language engineering*, *10*(3-4), 211–225. https://doi.org/https://doi.org/10.1017/S135132490400350X

Ide, N., & Romary, L. (2006). Representing Linguistic Corpora and Their Annotations. *LREC*, 225–228.

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlỳ, P. (2010). Fast syntactic searching in very large corpora for many languages. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 741–747.

[6] https://github.com/postgrespro/rum (September 24, 2023).

Krause, T., & Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, *31*(1), 118–139.

Leach, P., Mealling, M., & Salz, R. (2005). *A Universally Unique IDentifier (UUID) URN Namespace* (tech. rep.).

Lehmann, H. M., & Schneider, G. (2012). BNC Dependency Bank 1.0. In S. O. Ebeling, J. Ebeling, & H. Hasselgård (Eds.), *Studies in variation, contacts and change in English, volume 12: Aspects of corpus linguistics: Compilation, annotation, analysis*. Varieng. https://varieng.helsinki.fi/series/volumes/12/

Manning, C. D. (2009). *An Introduction to Information Retrieval*. Cambridge university press.

Meurers, W. D. (2005). On the use of electronic corpora for theoretical linguistics: Case studies from the syntax of German. *Lingua*, *115*(11), 1619–1639.

Rayson, P. E., Mariani, J. A., Anderson-Cooper, B., Baron, A., Gullick, D. S., Moore, A., & Wattam, S. (2017). Towards Interactive Multidimensional Visualisations for Corpus Linguistics. *Journal for language technology and computational linguistics*, *31*(1), 27–49.

Rohde, D. L. T. (2005). Tgrep2 user manual. http://tedlab.mit.edu/%5C~dr/Tgrep2/tgrep2.pdf

Schneider, R. (2013). Kogra-db: Using mapreduce for language corpora. In M. Horbach (Ed.), *Informatik 2013 – informatik angepasst an mensch, organisation und umwelt* (pp. 140–142). Gesellschaft für Informatik e.V.

Topics in Swedish News on Climate Change: A timeline 2016 – 2023

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

Abstract

The CLARIN node the "National Language Bank of Sweden" makes a corpus of Swedish television news available for research. We have here (i) extracted recurring topics related to climate change from this corpus, and (ii) studied how these topics vary over time by implementing a timeline visualisation of the topic occurrences. We extracted all texts from the Swedish television news corpus that were published between 2016 and 2023, and which contain any one of the two keywords "climate change" or "climate crisis". We applied topic modelling on this corpus, which resulted in 35 stable topics being extracted. We, thereafter, applied the timeline visualisation implemented for plotting the 35 topics. The topic modelling and the timeline visualisation provided us with an overview of topics we expected to find in the corpus, as well as with topics we did not beforehand anticipate to occur frequently in news on climate change.

1 Introduction

There are many possible sub-topics that could be reported in the news in relation to the general topic of climate change. We here investigate such sub-topics in Swedish news over the past few years. More specifically, we investigate what topics have been discussed in relation to climate change in one specific textual source, the web pages of *SVT Nyheter*, i.e., the news from the Swedish public service television company.

In addition to producing news programmes, SVT Nyheter also publishes short web-based news articles that typically each one of them summarises a televised news piece. For the work presented here, we used a corpus containing texts from SVT Nyheter that is made available through the CLARIN node the "National Language bank of Sweden" (Språkbanken Text)[1]. The aim of our work using this corpus was (i) to find examples of recurring topics in news on climate change, (ii) to study and visualise the occurrence of the topics over time, and (iii) to showcase a method for visualising the output of topic modelling applied to a text collection that includes a temporal dimension.

This is not the first study that uses (semi-)automatic text mining methods for investigating how environmental issues are discussed (Barkemeyer et al., 2017; Barkemeyer & Frank Figge, 2009; Fischer et al., 2017; Flottum, 2017; Stede et al., 2023; Tvinnereim & Fløttum, 2015). However, we believe that neither the exact methods we apply – nor this corpus – have been used for this task before.

2 Method

From the SVT Nyheter corpus we extracted all texts published from 2016 until February 2023, and that contained either the Swedish translation of the word "climate change" or the word "climate crisis" (i.e., "klimatförändring" or "klimatkris").

We, thereafter, applied the topic modelling tool Topics2Themes (Skeppstedt et al., 2018) on the sub-corpus extracted[2]. We provided the topic modelling algorithm with an extensive stop word list, as well as with a list of collocations, both of them manually compiled through iteratively running the topic modelling tool on the sub-corpus and inspecting the output. In the final iteration, we ran the topic modelling

[1]The corpus, which is collected from 2004, can be found here: https://spraakbanken.gu.se/resurser/svt
[2]Topics2Themes is a language-independent topic modelling tool: https://github.com/mariask2/topics2themes

algorithm 50 times, each time instructing it to return 50 topics. From these re-run outputs we kept the ten most typical ones, and only retained topics that occurred in all ten outputs.

Finally, we implemented and applied a timeline-based visualisation of the topic modelling output. The visualisation indicates each news article in the climate change sub-corpus by a vertical *text-line*, with its X-axis position determined by the date the article was published. For each topic extracted by the topic modelling algorithm, a horisontal *topic-line* is drawn on the Y-axis and labelled with the most typical words for the topic. If a topic is included in an article, this is indicated by thicker vertical bars on the part of the text-line that crosses the topic-line. If several climate change articles are published the same day, their vertical lines are positioned adjacent to each other on the X-axis.

3 Results and Discussion

When using our two search terms, a total of 2,270 news articles were extracted. Applying topic modelling on this sub-corpus resulted in 35 stable topics being generated. The timeline and top terms for the topics are shown in Figure 1. The first topic is generated from articles that seem to summarise a number of different television news pieces, and is not a coherent topic. There are also some very wide topics – e.g. topics about Sweden, Russia or about reports of statistics – in which climate change is not the only focus of the article, although it is mentioned. Most topics are, however, focused on climate change. All topics extracted are described in Table 1.

Some topics and their corresponding timeline are very predictable – e.g., that the news start reporting about Greta Thunberg in the end of 2018 – which gives some kind of indication of external validity to the method. There are also less expected topics – e.g., that the news regularly mentions the aspect of climate change when reporting about moose – which could serve as an indication of that this method could help us learn something new.

With this visualisation, we aim to generate a comprehensible overview of the topics and when they occur in the news. The visualisation also gives an overview of the general timeline of the articles of the sub-corpus. It can, for instance, be seen that in some time periods in 2020, there are fewer articles that mention climate change, possibly due to the Covid-19 pandemic dominating the news.

It should, however, be noted that it is not enough to study the topic timelines and the words extracted to represent this topic, but that examples of the texts typical for the topics also need to be studied in order to understand the topic. The graphical user interface of Topics2Themes supports such an exploration of texts typical to a topic, and we used this interface for reading a few texts for each topic extracted.

The programming code for generating the timelines is provided as open source[3]. We have previously generated topic modelling timelines that are somewhat similar to those produced here (Stede et al., 2023). However, these previous timeline visualisations were specifically constructed for the output of that study, while we here provide programming code that is possible to apply more generally to the topic modelling output of the Topics2Themes tool.

We have used the topic timeline code to generate similar timeline visualisations for several other types of corpora, and we will continue to evaluate its usefulness for supporting researchers in exploring text corpora. To make it possible to easily switch between close and distant reading, it might be useful if interesting areas on the timeline visualisation could form points of departure for exploring the content of the texts. It might, for instance, be relevant to make it possible to select a text for further exploration by clicking on the vertical line that represents it. Investigating such a timeline-based text exploration will therefore be the next step in our work.

Acknowledgments

The author would like to thank Paul Rosenbaum for input on the design of the visualisation.

The programming code for topic-based text visualisations was created as a part of InfraVis, the Swedish National Research Infrastructure for Data Visualization (the Swedish Research Council, grant

[3] At: https://github.com/CDHUppsala/topic-timelines. (Configuration files used for Topics2Themes are also provided there.)

2021-00181).

References

Barkemeyer, R., Figge, F., Hoepner, A., Holt, D., Kraak, J. M., & Yu, P.-S. (2017). Media coverage of climate change: An international comparison. *Environment and Planning C: Politics and Space*, *35*(6), 1029–1054. https://doi.org/10.1177/0263774X16680818

Barkemeyer, R., & Frank Figge, T. H. (2009). What the papers say: Trends in sustainability. a comparative analysis of 115 leading national newspapers worldwide. *The Journal of Corporate Citizenship*, *33*, 69–86.

Fischer, D., Haucke, F., & Sundermann, A. (2017). What does the media mean by 'sustainability' or 'sustainable development'? an empirical analysis of sustainability terminology in german newspapers over two decades. *Sustainable Development*, *25*(6), 610–624. https://doi.org/https://doi.org/10.1002/sd.1681

Flottum, K. (2017). *The role of language in the climate change debate* (First edition.). Taylor; Francis.

Skeppstedt, M., Kucher, K., Stede, M., & Kerren, A. (2018). Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics. *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, 9–16.

Stede, M., Bracke, Y., Borec, L., Kinkel, N. C., & Skeppstedt, M. (2023). Framing climate change in Nature and Science editorials: applications of supervised and unsupervised text categorization. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-023-00199-7

Tvinnereim, E., & Fløttum, K. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature climate change*, *5*(8), 744–747.

Nr	Description
1:	Incoherent topic
2:	General about carbon dioxide emissions, effects and reduction strategies
3:	The European Union, in relation to climate change and other issues
4:	Weather reports where climate change is mentioned
5:	Loss of biodiversity and invasive species due to climate change
6:	How to handle drinking water supply issues in Sweden
7:	Party politics, especially related to the Green party
8:	Lack of snow due to climate change
9:	Greta Thunberg and climate activism
10:	Sweden
11:	Publications of IPCC reports
12:	The effect of climate change on the Arctic region
13:	Lack of adaption to climate change within the municipalities in Sweden
14:	Forrests and forresty, affecting and affected by climate change
15:	Floods and heavy rain
16	Community-based collection of ticks to monitor changes due to climate change
17:	Climate negotiations, relations between rich and poor countries
18:	SMHI (Swedish Meteorological and Hydrological Institute)
19:	Opinion poll statistics (e.g. about climate anxiety) and other results
20:	Forrest fires, e.g., in California
21:	China in relation to climate change
22:	Russia
23:	The Baltic Sea, e.g. affected by climate change, fishing and algal bloom
24:	Community reports on signs of spring to monitor climate change effects
25:	Reports on scientific studies that in some way are related to climate change
26:	Increase in tick-borne encephalitis, e.g., due to climate change
27:	Lack of drinking water, draughts
28:	Donald Trump, often in relation to climate change
29:	The role of the Amazon in climate change
30:	Civil disobedience-based climate activism
31:	Museum exhibitions related to climate change
32:	Food, food choices and food production in relation to climate change
33:	Heat waves, particularly in Europe
34:	Moose and how they might be affected by climate change
35:	Negative effects on reindeer herding from climate change

Table 1: A manually authored description of each of the 35 topics extracted, after having read a few of the most typical news articles for each topic.

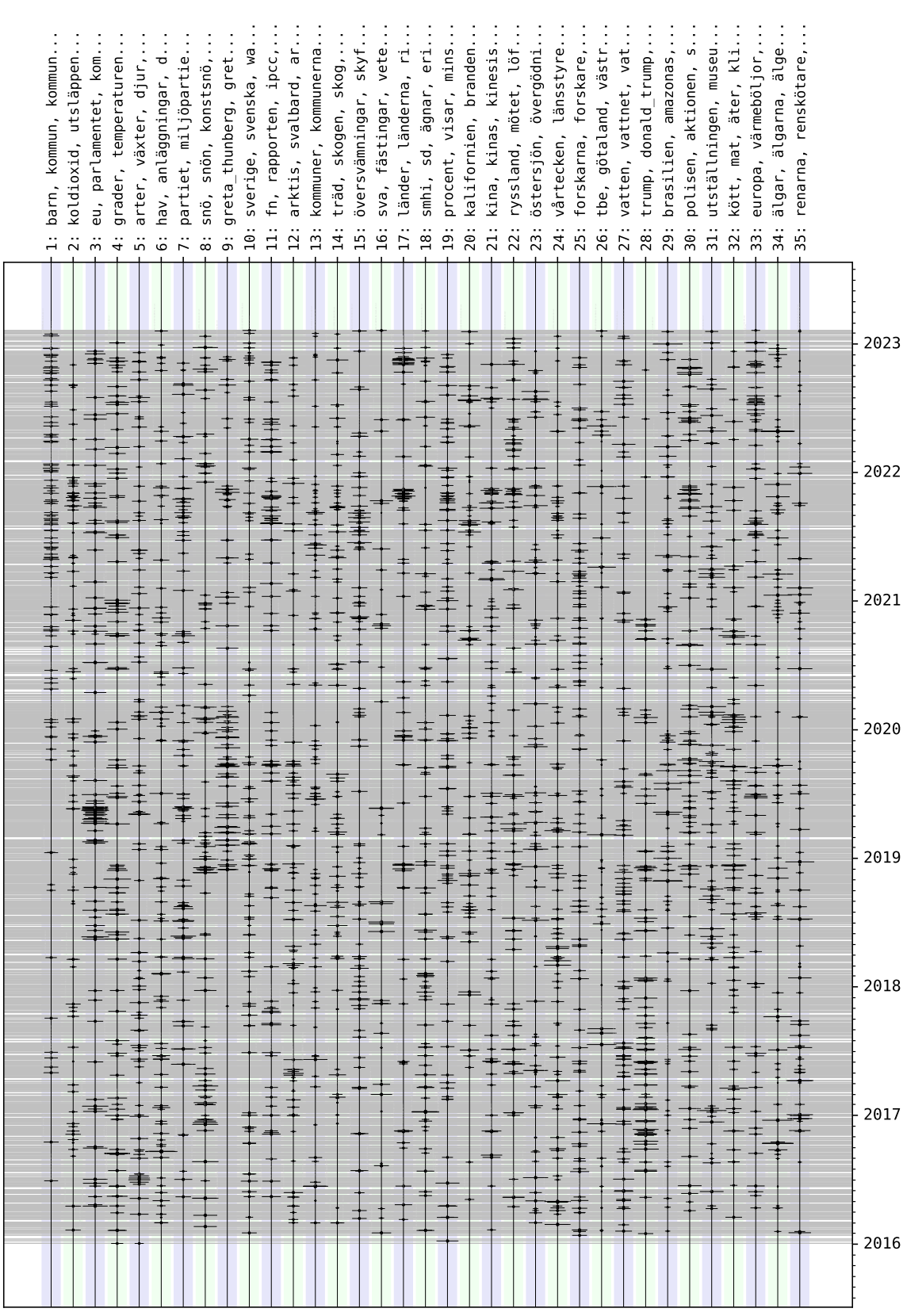


Figure 1: Timeline for the 35 topic modelling topics detected by the topic modelling algorithm. Each grey vertical line represents an article in the climate change sub-corpus. The black bars crossing the horisontal lines representing the topics indicate occurrences of a topic in the article.

DBBErt: Part-of-Speech Tagging of Pre-Modern Greek Text

Colin Swaelens and **Els Lefever**
Language Technology & Translation Team
Ghent University, Belgium
firstname.lastname@ugent.be

Ilse De Vos
Department of Linguistics
Ghent University, Belgium
i.devos@ugent.be

Abstract

This contribution presents DBBErt, a machine-learning approach to linguistic annotation for pre-Modern Greek, which provides a part-of-speech and fine-grained morphological analysis of Greek tokens. To this end, transformer-based language models were built on both pre-Modern and Modern Greek text and further fine-tuned on annotated treebanks. The experimental results look very promising on a gold standard of Byzantine book epigrams, with an F-score of 83% for coarse-grained part-of-speech-tagging and of 69% for fine-grained morphological analysis. The resulting pipeline and models will be added to the CLARIN infrastructure to stimulate further research in NLP for Ancient and Medieval Greek.

1 Introduction

The Database of Byzantine Book Epigrams or DBBE (Ricceri et al., 2023) contains over 12,000 epigrams. They are stored both as *occurrences* – the epigrams exactly as they occur in the manuscripts – and as *types* – their orthographically normalised counterparts. The relationship between occurrences and types is not one-to-one. For instance, Example 1 (DBBE Type 2148, translated by the authors) represents 70 two-verse occurrences of the ὥσπερ ξένοι epigram which was used widely by scribes to mark their joy of reaching the end of the manuscript and thus of their copying task.

The decision to link multiple occurrences to a single type was both pragmatic and conceptual. Creating fewer types not only freed up time to trace new occurrences, it was also a straightforward way to group similar occurrences. Soon however, this all-or-nothing system ran against its limitations: What exactly does "similar" mean? How "similar" do occurrences need to be for them to be put under the same type? The ὥσπερ ξένοι epigram for example circulated in many different versions, some counting three or four verses. To deal with this variety, increasingly more types were created, each of them covering different subsets of occurrences.

To (re)connect these subsets, a complementary system was introduced to link individual verses regardless of the type their occurrence belongs to. As for the ὥσπερ ξένοι epigram, no less than 202 instances of its first verse are to be found in DBBE. Although a huge step forward, this system still treats similarity as a dichotomy whereas it clearly is a continuum. Also, it does not allow to adequately visualise variation within more complex lists of "similar" verses nor to take into account different parameters, both textual and other. In order to add linguistic information enabling more advanced similarity detection and visualisation, we developed the first annotation tool for non-normalised Byzantine Greek.

(1) Ὥσπερ ξένοι χαίρουσιν ἰδεῖν πατρίδα,
οὕτως καὶ οἱ γράφοντες βιβλίου τέλος.
Just as strangers rejoice upon seeing their homeland,
so do writers upon completion of a book.

2 Related research

The last two decades witnessed a number of initiatives for compiling pre-Modern Greek corpora and making them accessible in open-source format, which stimulated NLP research on Ancient and Medieval Greek texts. An important corpus initiative is the Open Greek and Latin Project,[1] which consists of the Perseus Digital Library (G. R. Crane, 2022), a collection of more than 13,5M tokens of mostly classical Greek prose and poetry, and the First1K Project, containing 25,5M tokens of classical and post-classical Greek prose and poetry.[2]

In addition, various treebanks were developed. The Ancient Greek Dependency Treebank (AGDT) (Bamman & Crane, 2011) stores 560,000 tokens from both classical prose and poetry, all of which were tagged manually. The PROIEL (Haug & Jøhndal, 2008) treebank has a more specific content, as it stores 277,000 tokens of the New Testament in Greek and four other languages. The Gorman treebank (Gorman, 2020) on the other hand contains around 550,000 tokens of exclusively classical Greek prose. Finally, the Pedalion Trees (Keersmaekers et al., 2019) count around 320,000 tokens of annotated texts complementary to the AGDT, among which Trismegistos (Depauw & Gheldof, 2014), a database of papyri displaying the original text with all its idiosyncrasies including *errors*, just like the DBBE occurrences. All of the above mentioned treebanks are annotated in accordance with the Universal Dependencies principles and guidelines (Nivre et al., 2017).

A widely known and used tool for automatic linguistic annotation of Ancient Greek is Morpheus (G. Crane, 1991), a rule- and dictionary-based system that performs part-of-speech tagging and morphological analysis. It has, however, two important shortcomings: (1) it does not disambiguate ambiguous forms, but instead returns all possible analyses, and (2) it cannot analyse out-of-vocabulary words. In order to cope with this lack of flexibility, researchers have started to develop machine-learning systems for Greek part-of-speech tagging. Celano et al. (2016) did a comparative study, which showed that MateTagger (Bohnet & Nivre, 2012) outperformed Hunpos tagger (Halácsy et al., 2007), RFTagger (Schmid & Laws, 2008), the OpenNLP part-of-speech tagger[3] and NLTK Unigram tagger (Bird, 2006) on Ancient, normalised Greek data. Keersmaekers (2019), however, obtained different results when applying various taggers on papyrological data, with RFTagger clearly outperforming the other taggers on this specific data type. More recent approaches rely on neural networks, such as RNN tagger Schmid (2019), and the transformer-based part-of-speech tagger developed by Singh et al. (2021), which showed very promising results on an evaluation set containing normalised DBBE types.

3 Part-of-Speech Tagger

To develop a part-of-speech tagger for Ancient and Byzantine Greek, we compared three different transformer-based language models with embedding representations: BERT (Devlin et al., 2018), ELECTRA (Clark et al., 2020), and RoBERTa (Liu et al., 2019). These were then fine-tuned on the task of both coarse-grained part-of-speech tagging and fine-grained morphological analysis. To train these models, two data sets were compiled: one consisting of all Ancient and Byzantine Greek text corpora described in Section 2, and that same set complemented with the Modern Greek Wikipedia data. This allowed us to ascertain whether or not Modern Greek contributes to the modelling of Byzantine Greek.

For the supervised task of part-of-speech tagging and morphological analysis, we compiled a training set based on the treebanks described in Section 2 and completed it with a small set of 2,000 manually annotated tokens from DBBE occurrences. To train the part-of-speech tagger, we made use of the FLAIR framework (Akbik et al., 2019), where the contextual token embeddings from the language models are stacked with randomly initialised character embeddings. These are processed by a bi-directional long short-term memory encoder (hidden size of 256) and a

[1] https://opengreekandlatin.org
[2] https://opengreekandlatin.github.io/First1KGreek/
[3] https://opennlp.apache.org

conditional random field decoder. For evaluation, a gold standard containing 10,000 tokens of non-normalised Byzantine Greek epigrams out of the DBBE corpus was compiled, manually annotated and validated through an inter-annotator agreement study.

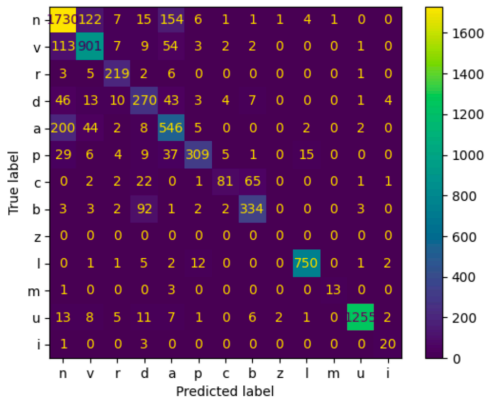


Figure 1: Confusion matrix of the coarse-grained part-of-speech labels

The BERT model trained on Classical, Medieval and Modern Greek performs best on this task with an F-score of 83% on the coarse-grained part-of-speech tagging and 69% on the morphological analysis. This model is hence called DBBErt.[4] Figure 1 shows the confusion matrix of the coarse-grained part-of-speech tagging, which reveals some expected trends (e.g., lot of confusion of the label noun (n) with the label adjective (a)).

4 Conclusion and Future Research

This paper introduced DBBErt, a transformer-based Part-of-Speech tagger for Ancient and Byzantine Greek. The evaluation of the tool on a novel gold standard containing occurrences of Byzantine Epigrams showed very promising results. In future research, we will keep improving DBBErt, since we believe that automatic linguistic annotation of non-normalised text will be very valuable for NLP research on historic languages. With respect to the DBBE, the enriching of the epigrams with linguistic information will supply important additional information for the next step in our research: measuring similarity between lexical variants of words in order to detect relationships between verses in various occurrences.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59.

Bamman, D., & Crane, G. (2011). The ancient greek and latin dependency treebanks. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language technology for cultural heritage* (pp. 79–98). Springer Berlin Heidelberg.

Bird, S. (2006). Nltk: The natural language toolkit. *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 69–72.

Bohnet, B., & Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455–1465.

[4]The model is made available at https://huggingface.co/colinswaelens

Celano, G. G. A., Crane, G., & Majidi, S. (2016). Part of speech tagging for ancient greek. *Open Linguistics*, *2*(1).

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, *6*(4), 243–245.

Crane, G. R. (2022). Perseus digital library [Last accessed 14 October 2022].

Depauw, M., & Gheldof, T. (2014). Trismegistos: An interdisciplinary platform for ancient world texts and related information. *Theory and Practice of Digital Libraries – TPDL 2013 Selected Workshops*, 40–52.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gorman, V. B. (2020). Dependency treebanks of ancient greek prose. *Journal of Open Humanities Data*, *6*(1).

Halácsy, P., Kornai, A., & Oravecz, C. (2007). Hunpos: An open source trigram tagger. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 209–212.

Haug, D. T., & Jøhndal, M. (2008). Creating a parallel treebank of the old indo-european bible translations. *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, 27–34.

Keersmaekers, A. (2019). Creating a richly annotated corpus of papyrological Greek: The possibilities of natural language processing approaches to a highly inflected historical language. *Digital Scholarship in the Humanities*, *35*(1), 67–82.

Keersmaekers, A., Mercelis, W., Swaelens, C., & Van Hal, T. (2019). Creating, enriching and valorizing treebanks of Ancient Greek. *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 109–117.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nivre, J., Zeman, D., Ginter, F., & Tyers, F. (2017). Universal Dependencies. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*.

Ricceri, R., Bentein, K., Bernard, F., Bronselaer, A., De Paermentier, E., De Potter, P., De Tré, G., De Vos, I., Deforche, M., Demoen, K., Lefever, E., Rouckhout, A.-S., & Swaelens, C. (2023). *The database of byzantine book epigrams project: Principles, challenges, opportunities* [working paper or preprint].

Schmid, H. (2019). Deep learning-based morphological taggers and lemmatizers for annotating historical texts. *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, 133–137.

Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, 777–784.

Singh, P., Rutten, G., & Lefever, E. (2021). A pilot study for bert language modelling and morphological analysis for ancient and medieval greek. *The 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, co-located with EMNLP 2021*, 128–137.

Emotion and Abstractness in Austrian Parliamentary Discourse

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

Klaus Hofmann
University of Vienna
Vienna, Austria
klaus.hofmann@univie.ac.at

Abstract

We investigate Austrian parliamentary discourse styles by combining utterances from the Corpus of Austrian Parliamentary Records with a large dataset of affective norms (Köper and Schulte im Walde, 2016). The results suggest that parliamentary discourse styles differ depending on gender, party affiliation and utterance type (regular speech vs. interjections). We also find evidence for a characteristically male right-wing populist mode of parliamentary interaction marked by negative and non-abstract language use.

1 Introduction and related research

Political discourse has proven a rich source for research in the humanities and social sciences and beyond. Parliamentary records are often at the centre of this research, as they rank among the most prototypical and best-documented types of political discourse. A growing number of machine-readable and annotated parliamentary corpora have become available in recent years (see e.g. Fišer, Lenardič & Erjavec, 2018: 1321; Ogrodniczuk et al., 2022), facilitating computer-based and specifically quantitative approaches. A number of initiatives and projects in the context of CLARIN are dedicated to the creation and analysis of parliamentary-based sources for example the ParlaCLARIN recommendations for encoding parliamentary records (Erjavec & Pančur, 2022) or the multilingual comparable ParlaMint data set (Erjavec et al., 2023a; Erjavec et al., 2023b).
The present study focuses on parliamentary discourse in Austria. The study is being conducted within the CLARIAH-AT context, which forms part of the larger CLARIN enterprise. The aim of the study is to explore how lexical sentiment data (i.e. the emotional value and strength of words) and abstractness ratings may inform our understanding of parliamentary discourse in Austria. Specifically, we ask (a) to what degree language usage as defined by these metrics is related to factors such as gender, party membership and parliamentary role, (b) to what degree language usage is subject to change over time, and (c) whether usage differs in different utterances types, i.e. regular speeches (1) vs. interjections (2):

(1) Abg. Dr. Josef Cap: […]. Hier in diesem Hause sitzen keine Idioten, und daher werden Sie hier immer wieder von uns vorgeführt werden für die Politik, für die Sie stehen. […]
['There are no idiots sitting here in this House, and therefore you will be exposed for the politics that you stand for again and again.']

(2) Abg. Dr. Petrovic: Das haben Sie verkürzt und falsch zitiert! ['You are quoting this in an abbreviated and misleading way!'] […]
Abg. Haigermoser: Wahnsinn! ['Insanity!'] […]

Sentiment analysis has been applied to parliamentary speeches before (Abercrombie & Batista-Navarro, 2020). However, few studies have integrated sentiment with abstractness scores in an attempt to profile

linguistic usage more broadly. Studies on parliamentary discourse in the Austrian context are generally still rare (but see Kern et al., 2021, Haselmayer et al., 2022).

2 Data and method of analysis

2.1 Corpus of Austrian Parliamentary Records

Shortly preceding the release of the ParlaMint 3.0 dataset (Erjavec et al., 2023b), which now also contains Austrian data, this study makes use of a somewhat older and smaller data set, namely the Corpus of Austrian Parliamentary Records, with slightly different mark-up. The corpus contains the parliamentary records of the National Chamber (*Nationalrat*), one of two chambers of the Austrian parliament from the XXth to the XXVth legislative periods, covering the years between 1996 and 2017. It is based on official transcripts (produced from shorthand) (Wissik & Pirker, 2018). Besides being tokenized, part-of-speech tagged and lemmatized, all speeches delivered by members of parliament (as well as verbal interjections) are annotated as utterances and each speaker is identified and marked up, accordingly. Thus, every utterance can be linked to a specific speaker. For each speaker additional metadata is provided, such as gender, party membership etc. Other information provided by the stenographers like applause, interruptions, descriptions of scenes or gestures, description of procedures etc. are annotated as notes. In its entirety, the corpus contains approximately 75 million tokens representing over 600 000 word forms and 400 000 lemmas.

2.2 Dictionary of German affective norms

The dictionary of German affective norms (Köper & Schulte im Walde, 2016) contains 350 000 German lemmas (including nouns, verbs, adjectives and adverbs), automatically rated via a supervised learning algorithm on four affective dimensions, namely abstractness/concreteness, arousal, imageability and valence. Abstractness/concreteness measures the degree to which the concepts denoted by the words are accessible to the human senses (e.g. *Ball* 'ball' vs. *Theorie* 'theory'). Arousal describes the intensity of emotion provoked by a lexical stimulus (e.g. *ruhig* 'calm' vs. *gewalttätig* 'violent'). Imageability refers to the degree to which concepts can be experienced through human vision (e.g. *Tisch* 'table' vs. *Glaube* 'belief'). Valence refers to the value of the emotional response elicited by a word, which can be positive or negative (e.g. *Geschenk* 'gift' vs. *Strafe* 'punishment').

One major advantage of using this data set over others is its size, which the compilers achieved through propagation from human-rated seed words using deep-learning-based skip-gram embeddings (also known as *word2vec*, Mikolov et al. 2013). We are aware that this data set is not specifically tailored to Austrian German, but in the absence of a comparable and similarly comprehensive data set for Austrian usage we still opted for it in the interest of maximizing coverage.

2.3 Data processing and method of analysis

A subcorpus including utterances by members of parliament (speeches and interjections) but excluding all procedural content was compiled from the parliamentary corpus. All speaker variables, including speaker gender, party membership and parliamentary role (government vs. opposition), were linked directly to the utterances. A custom stopword list excluded function words, titles, recurrent phrases (e.g. *Bundeskanzler*, *Hohes Haus*) and party names. The resultant utterance subcorpus was merged with the affective norms data set.

We opt for linear regression modelling to analyse the data set. Arousal, valence, concreteness and imageability serve as dependent variables. The predictor variables include all collected speaker-related information, as well as utterance type (speech vs. interjection) and legislative period. Compared to more sophisticated deep learning approaches, linear regression is a relatively simple but transparent and accessible method for investigating the relative impacts of a range of predictors on the dependent variables. All calculations are carried out with R (R Core Team, 2023).

3 Results and discussion

Our results point to marked differences in language use across all variables under investigation. In line with previous research on gendered language usage, female members of the Austrian parliament generally display more positive (Danner et al., 2001; Mehl & Pennebaker, 2003) and more arousing (Thomas

& Murachver, 2001) language compared to their male colleagues. Contrary to expectation, however, men display higher concreteness scores than women. This does not straightforwardly match previous findings that men utilize abstract communication as a signal of status and power (Wakslak et al., 2014; Joshi et al., 2021) or that women discuss policies in more concrete terms (Hargrave & Langengen, 2021). Emotion and abstractness also vary significantly with party affiliation, such that right-wing parties (*FPÖ* and its short-lived spin-off *BZÖ*) use concrete language to a larger degree than liberal parties (*GRÜNE*, *NEOS*), while the traditional center parties (*ÖVP*, *SPÖ*) are located in the middle of this cline. Additionally, the right-wing parties set themselves apart by more negative language compared to the other parties.

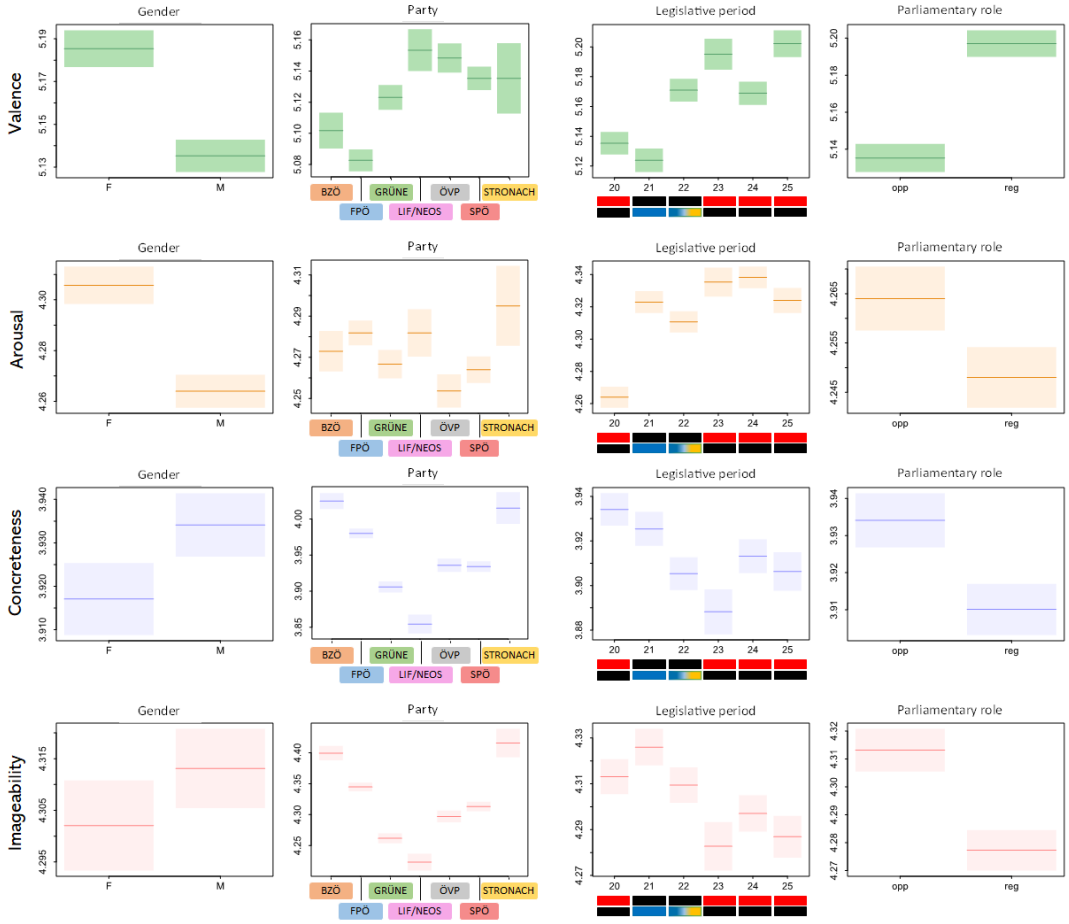


Figure 1: Results of the linear regression analysis for valence, arousal, concreteness and imageability.

We interpret these findings in terms of a characteristically male right-wing populist mode of parliamentary discourse, characterized by negative and non-abstract lexis, in distinction to liberal and establishment styles. Seen in this light, lack of abstract language in parliamentary debates serves to signal anti-establishmentism rather than lack of power.

Language use also correlates strongly with the parliamentary roles of government and opposition. Members of the opposition display lower valence, higher arousal, and higher concreteness/imageability scores, which aligns well with the two complementary functions in parliament. Diachronically, parliamentary discourse generally developed towards more positive and more aroused language, while abstractness trended towards the more concrete and imageable end of the spectrum.

Considering the distribution of interjections vis-à-vis regular speeches, our analysis reveals that male members of parliament and members of right-wing parties display a much higher likelihood to engage in these unsolicited communicative strategies compared to women and liberal or establishment parties. This further adds to the picture of a male-dominant populist mode of parliamentary expression.

4 Conclusion and future work

Our study delineates some of the major characteristics in Austrian parliamentary discourse styles, particularly in terms of how they relate to membership demographics, party affiliation and parliamentary roles. The new version of the ParlaMint data, Parlamint 3.0 (Erjavec et al. 2023a), which also contains Austrian data, opens up new opportunities for comparative and contrastive studies with parliamentary corpora from other countries where sentiment and/or abstractness norms are available for the respective languages and can generate further insights into parliamentary discourse, similarities and country-specific details on the dimensions of emotion and abstractness. Finally, our macroscopic perspective invites complementary micro-level research using established qualitative methods of discourse analysis, which in turn may serve to inform and refine the quantitative approach presented here.

References

Abercrombie, G. & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science* 3, 245–270. https://doi.org/10.1007/s42001-019-00060-w

Danner, D. D., Snowdon, D. A., & Friesen, W. V. (2001). Positive emotions in early life and longevity: Findings from the nun study. *Journal of Personality & Social Psychology* 80, 804–813.

Erjavec, T., et al. (2023a). The ParlaMint corpora of parliamentary proceedings. *Language Resources & Evaluation*, 57, 415–448, https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T., et al. (2023b), *Multilingual comparable corpora of parliamentary debates ParlaMint 3.0*, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1486

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

Fišer, D., Lenardič, J. & Erjavec, T. (2018). CLARIN's Key Resource Families. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1320–1325. http://www.lrec-conf.org/proceedings/lrec2018/summaries/829.html

Hargrave, L., & Langengen, T. (2021). The gendered debate: Do men and women communicate differently in the House of Commons? *Politics & Gender* 17(4), 580–606. https://doi.org/10.1017/S1743923X20000100

Haselmayer, M., Dingler, S. C. & Jenny, M. (2022). How women shape negativity in parliamentary speeches: A sentiment analysis of debates in the Austrian Parliament. *Parliamentary Affairs* 75(4), 867–886. https://doi.org/10.1093/pa/gsab045

Joshi, P. D., Wakslak, C. J., Huang, L. & Appel, G. (2021). Genderd differences in communicative abstraction and their organizational implications. *Rutgers Business Review* 6(2), 145–153. https://ssrn.com/abstract=3965803

Kern, B. M. J., Hofmann, K., Baumann, A. & Wissik, T. (2021). Komparative Zeitreihenanalyse der lexikalischen Stabilität und Emotion in österreichischen Korpusdaten. In Katsikadeli, C., Sellner, M. & Gassner, M. (eds.). *Digital lexis and beyond*. *Selected papers from the 45th Austrian Linguistics Conference*, 104–118. https://epub.oeaw.ac.at/?arp=0x003c6407

Köper, M. & Schulte im Walde, S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016)*, 2595–2598. https://aclanthology.org/L16-1413

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality & Social Psychology* 84, 857–870.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C.J.C. Burges et al. (eds.). *Advances in Neural Information Processing Systems* 26, 3111–3119. https://doi.org/10.48550/arXiv.1310.4546

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, C., Matyáš Kopp, M. & Meden, K. (2022). ParlaMint II: The show must go on. *Proceedings of the Workshop ParlaCLARIN III @ LREC2022*, 1–6. https://aclanthology.org/2022.parlaclarin-1.1

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Thomson, R., & Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology* 40, 193–208.

Wakslak, C. J., Smith, P. K., & Han, A. (2014). Using abstract language signals power. *Journal of Personality and Social Psychology* 107(1), 41–55. https://doi.org/10.1037/a0036626

Wissik, T. & Pirker, H. (2018). ParlAT beta Corpus of Austrian Parliamentary Records. *Proceedings of the Workshop ParlaCLARIN: Workshop on Creating and Using Parliamentary Corpora @ LREC2018*, 20–23.

Libraries as Data Infrastructures

Martin Wynne
University of Oxford, United Kingdom
martin.wynne@ling-phil.ox.ac.uk

Andreas Witt
Leibniz Institute for the German Language, Germany
witt@ids-mannheim.de

Peter Leinen
German National Library, Germany
P.Leinen@dnb.de

Sally Chambers
DARIAH-EU, Ghent Centre for Digital Humanities, Ghent University and KBR, Royal Library of Belgium
sally.chambers@dariah.eu

Abstract

The CLARIN and DARIAH European research infrastructures have a long history of collaboration and cooperation. One recent joint initiative has been to strengthen and deepen collaboration with national and major research libraries, with a particular focus on ways to facilitate the wider use of the extensive and culturally important digital datasets curated by libraries as research data. In order to further this goal, a series of workshops has been initiated, and a Conference of European National Librarians (CENL) Dialogue Forum has been established. Ongoing collaborative work includes a survey of existing collaborations between libraries and research infrastructures, an investigation of the potential for the creation of unique language models from digital library collections and an exploration of emerging initiatives such as the common European Data Space for Cultural Heritage.

1 Introduction

National Libraries have not only been pioneers in the development of data infrastructures, but they also play an essential role in facilitating research in the arts and humanities. Likewise, the continual growth of digital (digitised and born-digital) cultural heritage is crucial for arts and humanities researchers, especially for analysis and interpretation using digital methods (Tasovac et al, 2020). The digital data infrastructure landscape is currently in considerable flux, both nationally[1] and internationally[2]. Existing Research Infrastructures, such as DARIAH and CLARIN, are joining forces to contribute to the European Open Science Cloud (EOSC), for example through the establishment of the Social Sciences and Humanities (SSH) Open Marketplace[3]. In the cultural heritage space, emerging initiatives such as the common European Data Space for Cultural Heritage[4] and the European Collaborative Cloud for Cultural Heritage[5] are set to disrupt this landscape further, providing both challenges, as well as unprecedented opportunities for both libraries and research infrastructures alike. It is within this evolving context that the idea of a CENL Dialogue Forum on Libraries as Data Infrastructures was born.

[1] ESFRI National Roadmaps [https://www.esfri.eu/national-roadmaps]
[2] Strategy Report on Research Infrastructures Roadmap 2021 [https://roadmap2021.esfri.eu/]
[3] Social Sciences and Humanities Open Marketplace [https://marketplace.sshopencloud.eu/]
[4] Common European Data Space for Cultural Heritage [https://digital-strategy.ec.europa.eu/en/news/deployment-common-european-data-space-cultural-heritage]
[5] Collaborative Cloud for Cultural Heritage [https://ec.europa.eu/commission/presscorner/detail/en/IP_22_3855]

2 Conference of European National Librarians

CENL, the Conference of European National Librarians[6], brings together the National Libraries of Europe. It is a network of 46 national libraries in 45 European countries in the Council of Europe. Founded in 1987, the mission of CENL is to advance the cause of Europe's national libraries through collaboration to preserve the continent's cultural heritage and make it accessible to all, with a specific focus on skills and knowledge exchange. Collaboration between libraries and research infrastructures such as DARIAH and CLARIN is not new. As well as an active CLARIN and Libraries community, which holds regular workshops, DARIAH has been exploring the inter-relationship between digital collections and digital scholarship together with library organisations such as LIBER, Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries[7] and IFLA, International Federation of Library Associations and Institutions[8], and is an active participant in the International GLAM Labs Community[9].

To facilitate structural and strategic collaboration between Europe's National Libraries and Research Infrastructures, the idea of a CENL Dialogue Forum was born. It provides an ideal opportunity to assess the landscape; identify and prioritise specific challenges and opportunities, and understand how (national) libraries could benefit from structural collaboration with, and active participation in Research Infrastructures such as DARIAH and CLARIN. A key issue for debate is the international accessibility of FAIR (Findable, Accessible, Interoperable and Reusable) datasets and related challenges in implementation. Furthermore, the Collections as Data initiative is gaining traction internationally[10]. With the increasing emergence of 'data labs' throughout the library community, such labs could be an ideal point of intersection between the libraries, research infrastructures and digital humanities research communities. Not only could the Dialogue Forum be the voice of libraries in this data space, at the same time, it would raise awareness of this crucial topic throughout the (national) library community.

A survey of national libraries was carried out from May to July 2023 with the aim of obtaining a deeper understanding of existing and planned collaboration with research infrastructures, and to elicit information about activities in the areas of digital scholarship and digital data curation. Questions covered areas including compliance with the FAIR principles and Open Science, collections as data, data labs, data access, digital literacy, artificial intelligence and data science.

Thirty-one National Libraries responded to the survey, of which 20 (64%) institutions indicated that "Participating in Research Infrastructure" is a strategic priority of the National Library. Overall, the responses indicate that the percentage of institutions actively engaged in research infrastructure, or intending to do so, is 82%. There are 23 (74%) responding institutions already active in National Research Infrastructure initiatives.

Of the participating institutions, 14 (45%) state that they are participating in a European initiative, and a further 6 institutions are planning to do so. Thus, at the European level, Research Infrastructure appears to be a current topic for more than half of the participating institutions. DARIAH (10) and CLARIN (9) are the most frequently mentioned initiatives. When asked about the different stages of development of individual topics, the institutions responded as shown in Figure 1.

[6] Conference of European National Librarians (CENL) [https://www.cenl.org/]
[7] Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries (LIBER) [https://libereurope.eu/]
[8] International Federation of Library Associations and Institutions (IFLA) [https://www.ifla.org/]
[9] International GLAM Labs Community [https://glamlabs.io/]
[10] Collections as Data: State of the Field [https://collectionsasdata.github.io/part2whole/iac/]

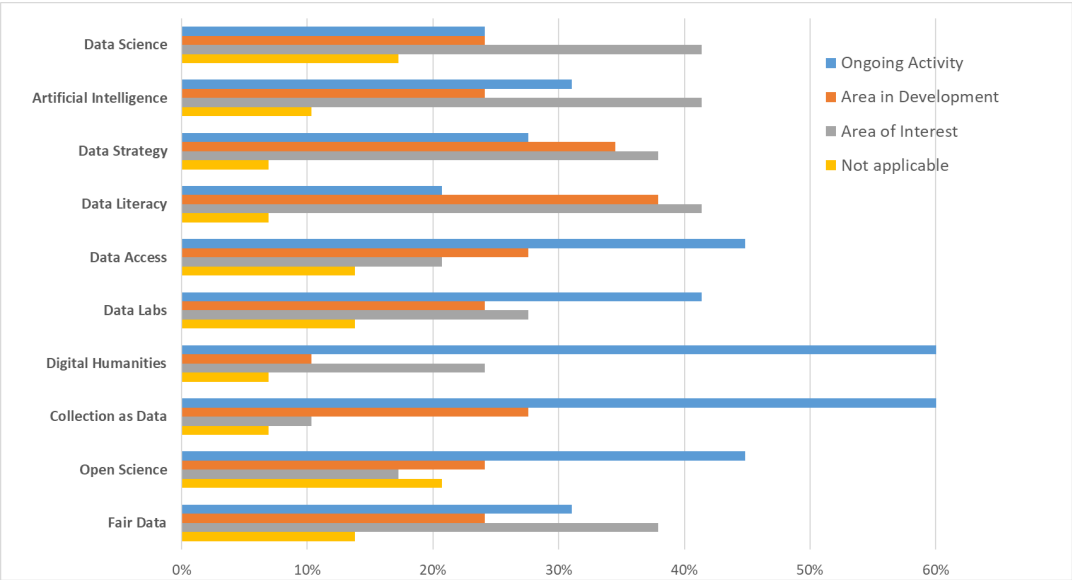


Figure 1: Results of the CENL Survey

3 Alignment of infrastructure projects

Together, the CENL Dialogue Forum, the ongoing series of CLARIN workshops and DARIAH's contribution to the common European Data Space for Cultural Heritage, will provide a platform for cooperation and collaboration, but will not directly achieve results in achieving interoperability and enhanced services for researchers. However, achieving interoperability and enhanced services for researchers will take place in libraries and national infrastructures, such as within specific projects:.

Text+[11] is a major new National Research Data infrastructure in Germany, involving a range of partners from academia, including CLARIN and DARIAH, and national and state libraries. The aim is to build a research data infrastructure focused on language and textual data, for a wide range of disciplines in the humanities and social sciences. The data which Text+ aims to deliver includes not only collections of historical texts, but also contemporary language corpora, lexical resources, and digital editions. Text+ will offer a comprehensive support infrastructure for all issues regarding collections, including interfaces, standards, authority data, and long-term preservation, etc.

DATA-KBR-BE[12] is a project at KBR, Royal Library of Belgium, which is developing an open data platform to offer data-level access to KBR's digitised and born-digital collections for digital humanities research. The project collaborates closely with the DARIAH and CLARIN consortia in Belgium, and builds on much recent and ongoing work in the area of 'collections as data'.

SSHOC-NL[13] is the latest in a series of joint CLARIN-DARIAH projects in the Netherlands, which will include the national library and national research institutes and which will build enhanced services for researchers, partly built on important past initiatives and collections such as Nederlab, Delpher and KB Lab Datasets.

Unlocking Digital Texts[14] is a collaboration between the Universities of Oxford, Cambridge and Notre Dame, with links to Text+ and Nederlab, which aims to make it easier to use a variety of textual

[11] Text+ [https://www.text-plus.org/en/home/]
[12] DATA-KBR-BE [https://www.kbr.be/en/projects/data-kbr-be/]
[13] SSHOC-NL Infrastructure awarded [https://www.huygens.knaw.nl/en/sshoc-nl-infrastructure-is-awarded-15-2-million-euros/]
[14] Unlocking Digital Texts [https://www.cdh.cam.ac.uk/research/projects/unlocking-digital-texts/]

formats as data in research. It will develop prototypes, and proofs-of-concept, building on existing standards (e.g. IIIF) and technologies rather than creating new formats or specific code dependencies.

The text digitization programme at the National Library of Norway has already created one of the largest text collections in the world and operates a DH-LAB that offers corpus services via a REST API and are also experimenting with the creation of language models based on the collections. Similarly, the National Library of France (BnF) Data Lab[15] is a service for researchers who wish to work with the BnF's digital collections.

Furthermore, there is the opportunity to align ongoing development in CLARIN online interfaces such as Korp, KonText, Corpuscle, NoSketchEngine etc. to more easily include library texts as datasets. Future development of the Virtual Language Language Observatory, Language Resources Switchboard and Federated Content Search could be optimised to work with more library collections and APIs.

4 Relevance to CLARIN

While research infrastructures for the arts humanities such as CLARIN and DARIAH have emerged in recent decades, for many centuries libraries have been the most important resource for researchers, and remain so in the digital age. For virtual, digital, distributed research infrastructures to be effective, they need to work closely with libraries, which play key roles as creators and curators of digital data, and as intermediaries between researchers and digital data, tools and expertise.

Creating language models from trusted and high-quality datasets is becoming an important area. Libraries not only offer access to large amounts of published material of known provenance and quality, but also unique opportunities to work with historical datasets, and thereby to create language models for a wider range of historical language varieties than is usually the case with existing research in the artificial intelligence, machine learning and natural language processing domains.

The ongoing collaborations will also provide an opportunity for libraries and research infrastructures to share knowledge and expertise, and potentially to share technical development work when it comes to user interfaces and APIs for sharing and using large text collections. Such collaborations should help to work against the inherent tendencies to waste effort, reinvent the wheel and create digital silos, and could become an important driver in the development and adoption of standards, common technological solutions and the interoperability of data collections and tools.

5 Conclusion

The initiative presented in this paper is intended to be an ongoing strategic collaboration, rather than a time-limited project, and will therefore inevitably be a work in progress rather than a completed discrete research activity. By the end of the summer of 2023, a number of activities currently underway, such as the CENL survey, will be completed, and others, such as the plan for the next CLARIN workshop, will be more firmly developed. However, given that a key aim is to include more participants and to promote openness and interoperability, it is important to disseminate the ongoing outputs as they happen, such as . promoting fruitful dialogue at the CLARIN Annual Conference.

References

Tasovac, T. A., Chambers, S., Tóth-Czifra, S. (2020). *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*. ⟨hal-02961317⟩

[15] BnF Data Lab [https://www.bnf.fr/fr/bnf-datalab]

A Multilingual Database for Icelandic L2 Flashcards

Xindan Xu
University of Iceland, Iceland
xindanxu@hi.is

Þórunn Arnardóttir
University of Iceland, Iceland
thar@hi.is

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

Abstract

The IceFlash 4K database is a newly developed multilingual resource for Icelandic vocabulary learning as a second language. It currently contains the 4,000 most frequently used words in Icelandic, with translations in four languages: English, Polish, Chinese and Ukrainian. The IceFlash 4K resource, including the flashcards and developer-friendly raw materials, is published under a CC BY 4.0 license. IceFlash 4K will help learners to learn Icelandic vocabulary more efficiently and it is a useful resource for teachers and language resource developers.

1 Introduction

Studies have shown that intentional learning methods, such as using flashcards for vocabulary learning, are very efficient and can produce great retention of knowledge. In this paper, we present our newly compiled multilingual database for developing flashcards for learning Icelandic, IceFlash 4K. It currently includes the 4,000 most common Icelandic words and their translations into four languages: English, Polish, Chinese and Ukrainian.

This database was used to produce flashcards for Icelandic L2 studies and is currently available in four languages: Icelandic–English, Icelandic–Polish, Icelandic–Chinese and Icelandic–Ukrainian. The original proposal, which was discussed in Xu and Ingason (2021), only included Icelandic–English. Whilst more language versions can be added to the database in the future, these languages were prioritized due to their importance within the Icelandic language learning environment. People of Polish origin comprise the largest immigrant group in Iceland (Hagstofa Íslands, 2021) while people of Ukrainian origin have been immigrating to other countries, following the Russian invasion into Ukraine in February 2022, including to Iceland, so it is important to provide both these groups with language aid to ease their adaptation to the Icelandic society. English is an international communication language and Chinese is one of the most widely spoken languages in the world, and the language which has the greatest number of native speakers. The languages chosen cover a large group of people who do not have access to learning materials. The IceFlash 4K database, along with the produced flashcards, is published under a CC BY 4.0 license and is available at the Icelandic CLARIN repository (Xu et al., 2023) and on GitHub.[1] The flashcards are available as a printable PDF version and a digital Anki version.[2] Currently, the English version of the Anki flashcards has been downloaded more than 1,000 times.

The paper is structured as follows. Section 2 discusses relevant research and resources. Section 3 describes the database and how it was created, while Section 4 details the evaluation process. Finally, we conclude in Section 5.

2 Theoretical background

Vocabulary learning is an important aspect in second language acquisition, as well as when acquiring a native language. Nation (2001, p. 60) found that the vocabulary size of native adults who have English as

[1] https://github.com/antonkarl/iceFlash4K

[2] See https://docs.ankiweb.net/background.html.

their native language is approximately 20,000 words, and that this size is very difficult to attain for those who are learning the language as a second language. He claims that one practical approach would be to learn the most common words first in the language. He believed that with good knowledge of the first 1,000 most common words, individuals could achieve a good understanding of about 81% of the written language and 85% of the spoken language. A knowledge of the first 4,000 words allows individuals to understand about 98% of the written language and 96% of the spoken language. Furthermore, the 98% scope is considered optimal for students to understand a second language without further assistance (Nation, 2001, p. 79).

Unlike vocabulary acquisition by native language learners, which is a gradual process of building up their vocabulary from language acquisition, learning a second language requires the same process but over a much shorter period of time in order to become proficient in a new language. Additionally, a long-term retention of the vocabulary knowledge is needed so that it can be called out immediately if necessary. According to Ebbinghaus (1913), our memory weakens over time, and it happens dramatically after being introduced to new learning materials. He conducted an experiment on himself to see how memory works and why people forget. He explained the process of forgetting by means of a line of data points which later became “Ebbinghaus’s forgetting curve". Other researchers have also experimented with different models to interpret forgetting (see e.g. Murre and Dros (2015)). Ebbinghaus (1913) described that when revisions are repeated and spread over a number of time intervals, the rate of forgetting can be delayed or slowed. This phenomenon is known as spacing effect.

Programs specially designed for digital flashcards generally utilize algorithms based on the concept of the forgetting curve and the spacing effect. One such program is called Anki (Damien Elmes, 2023), in which users can make their own multimedia flashcards. Anki flashcards consist of a question (reviewing material) on the front side and an answer (material to be learned) on the back side. During the reviewing process, both active recall testing and spaced repetition are utilized in the program, so that the learning materials are reviewed with different time intervals based on how well the user has learned them. Research (Nakata, 2016; Nation & Hunston, 2013; Webb et al., 2020) suggests that flashcards, especially when applied with spaced repetition, can be very efficient in enhancing vocabulary learning and creating long-term retention of the vocabulary knowledge.

3 A multilingual database

The IceFlash 4K database was designed for producing flashcards for Icelandic L2 studies, consisting of information that is most useful for Icelandic L2 learning. A word frequency list was collected from the Tagged Icelandic Corpus (MÍM; Helgadóttir et al., 2012) and the 4,000 most frequently used words in the corpus were selected as a base vocabulary list for using in the flashcards. The main database is in the form of four tsv files, one for each language version. Each line in the tsv files contains a word and a variety of information about it, including the word category, its frequency in the corpus, a sample sentence which shows the word’s usage in context, the phonetic transcription and the name of the corresponding audio file (which is included with the Anki flashcards), a translation of the most common meaning of the word and selected inflectional forms.[3] Figure 1 shows an example of an Icelandic–English flashcard, both in the PDF version and the desktop version of Anki. The flashcard shows the noun *ár* ’year’ with various pieces of information, including an example sentence and inflectional forms.

Various similar decks exist on AnkiWeb for other languages, but there are always some differences. For example, the most popular deck for French is quite similar[4] but for that deck, only some of the words include sample sentences and the information about inflection is somewhat less detailed.

After the original publication (Xu & Ingason, 2021), in which only an English translation was available, three additional languages were added to the collection: Polish, Chinese and Ukrainian. Furthermore, internal and external reviews have been carried out to evaluate both the content and functionality of the flashcards and the full database has been released using the infrastructure of CLARIN. Results

[3] Note that audio transcriptions of the vocabulary items are not available on the project GitHub page due to the size of the files.

[4] See the French deck on Ankiweb here: https://ankiweb.net/shared/info/893324022.

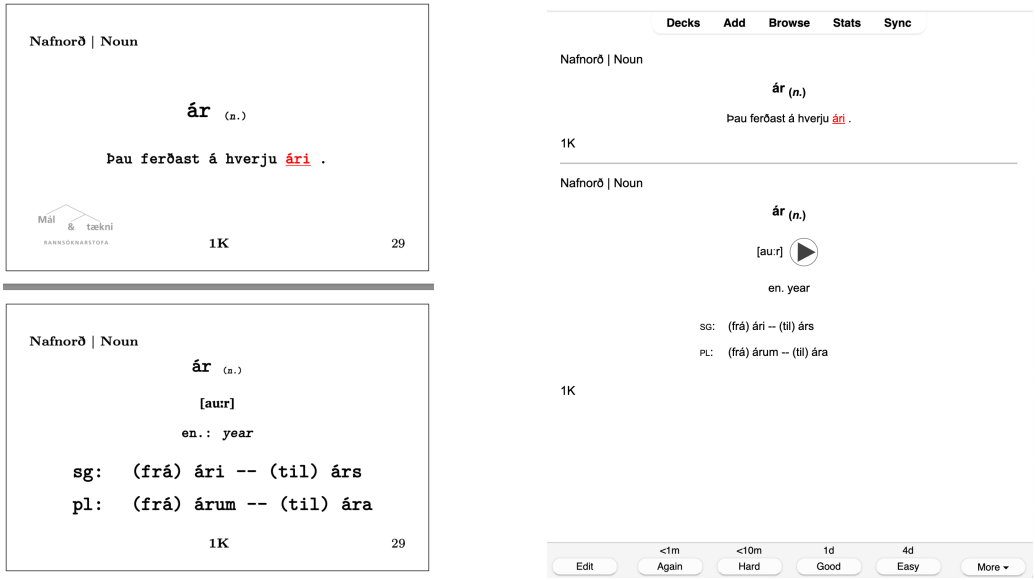


Figure 1: Flashcards for the noun *ár* 'year' in the PDF version (left) and the desktop version of Anki (right).

from the evaluations will be discussed further in Section 4.

The flashcards are created in nine main steps, apart from further improvements after the evaluations. The majority of the process was carried out automatically. See Xu and Ingason (2021) for a detailed account of data collection, processing and production of the flashcards.

The current flashcards do not contain visual stimuli. It is relatively difficult to find or produce images of certain word categories such as demonstrative pronouns, reflexive pronouns, certain verbs, and abstract nouns. Nevertheless, it would be helpful to add visual stimuli for some cards in future implementations.

4 Initial evaluation

Before the database and the flashcards were released to the public, internal evaluations and external user testing and interviews were carried out to ensure the quality of the resources.

4.1 Evaluation of phonetic transcription, translation and sample sentences

The phonetic transcriptions were originally generated automatically using the g2p-lstm model (Nikulásdóttir, 2020). Upon initial inspection, a few errors were found in the model's output, particularly when it comes to certain pronunciation rules in Icelandic, such as aspiration, preaspiration and *t*-insertion between consonant clusters, in combination with prefixed or compound words. As a result, three resources were combined to produce the final phonetic transcription of the words in the project: the Icelandic Pronunciation Dictionary (Rögnvaldsson, 2015), the Icelandic Pronunciation Dictionary for language technology (Nikulásdóttir, 2021) and an automatic transcription model (Nikulásdóttir, 2020). A total of 526 words were manually reviewed and corrected during this process, with reference to Rögnvaldsson (2020).

Translations of the Icelandic words were carried out in two steps. First, the list of words was automatically translated through the Google Translate web service. The resulting translation accuracy was poor in some cases, see Xu and Ingason (2021) for details. As a result, manual review and translation was carried out on all of the language versions, while translations from Google Translate as well as sample sentences were used as a reference. For the English version, translations were reviewed and corrected manually

by the authors and a proofreader. For the Polish and Chinese version, translations from Icelandic were done manually by native speakers, while for the Ukrainian version, a native speaker of Ukrainian, who speaks both English and Polish, manually translated the Icelandic words using English and Polish as intermediate languages.[5]

The sample sentences were collected from the Tagged Icelandic Corpus (MÍM), in which the majority of texts are from published books and online news, comprising approximately 50% of all the texts in the corpus (Helgadóttir et al., 2012). As a result, some sample sentences were considered too difficult for the students who are in the beginner levels of Icelandic studies, especially for the most common words in the first 1,000 tier. This was observed by the teachers of Icelandic as a Second Language in the University of Iceland. In response to this, sample sentences for the first 700 words were completely recreated using short sentences with easily understood vocabulary. This was carried out by a native speaker of Icelandic.

4.2 User testing of Anki flashcards

User testing of the Anki flashcards was carried out prior to the official release of the flashcards. A mini version of the Icelandic–English flashcards was created for this purpose, which consists of 200 words, with a random 5% of each 1,000 words tier. The mini version was consequently sent to volunteered participants, who were asked to use the test flashcards in Anki continuously for two weeks. After two weeks, the participants were asked to report back on their experience using the flashcards through an online form with 5-points likert scale items (Likert, 1932). The online form was completely anonymous and it was not possible to track answers back to individual participants. Apart from some background information, such as age group, native language, length of study of the Icelandic language etc., the form consists of 5 items about the functionality of Anki flashcards and 4 items about participants' Icelandic learning experience.

Survey questions	Strongly disagree (%)	Disagree (%)	Neutral (%)	Agree (%)	Strongly agree (%)
1. I find Icelandic easy to learn.	20	40	30	10	0
2. Knowing Icelandic is important to me in my personal life.	10	10	20	20	40
3. Knowing Icelandic is important to me in my work environment.	10	10	30	10	40
4. I enjoy learning languages through smart devices.	0	0	20	30	50
5. It was very easy to set up the flashcards on my smart device.	0	10	10	40	40
6. I find the flashcards very useful for learning new Icelandic vocabulary.	0	0	40	20	40
7. I find it very useful to have the audio transcriptions of the words.	0	10	0	20	70
8. I find it very useful to have the phonetic transcriptions of the words.	0	10	20	10	60
9. I find the inflectional forms very helpful.	0	11	0	33	56

Table 1: Feedback from initial user testing of Anki flashcards.

A total of 10 volunteered participants finished both testing and reporting on the feedback. Although this is not a big feedback dataset, it gives a general idea of the functionality of the flashcards that we created as well as recommendations on future improvements. Results from this testing is shown in Table 1. The majority of participants gave positive feedback on the different aspects of functionality of the flashcards (see items 5–9). For the Icelandic learning experience, the majority of the participants (60%) considers Icelandic not to be easy to learn (see item 1) and 60% considers that knowing Icelandic is important in their personal life (see item 2). Furthermore, we have received some anecdotal feedback from users who have had a positive experience using the resource.

5 Conclusion

We have presented the IceFlash 4K database, a newly developed multilingual resource for Icelandic vocabulary learning as a second language. It currently contains the 4,000 most frequently used words in Icelandic, with translations in four languages: English, Polish, Chinese and Ukrainian. By learning the

[5]Note that the authors are aware of the disadvantage of this method and an external evaluation will be carried out on the Ukrainian translation of the word list.

high frequency words, learners can understand a large portion of common texts such as newspapers and books. Furthermore, the words were put into four tiers, each containing one thousand words. We believe that this encourages learners to carry on learning and feel the sense of accomplishment when they finish the respective tier. Last but not least, both printable and digital flashcards were made so that learners can choose which format suits them best.

The database is available at the Icelandic CLARIN repository (Xu et al., 2023) under a CC BY 4.0 license. The importance of the currently described database not only lies upon its multilingual application of flashcards, but also its possibility for further language development and applications in other fields of study.

Acknowledgments

We thank the Áslaug Hafliðadóttir Memorial Fund for partially supporting our project, as well as the anonymous reviewers for their comments.

References

Damien Elmes, A. H., AMBOSS MD Inc. (2023, April 5). *Anki* (Version 2.1.57). https://apps.ankiweb.net

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. Teachers College Press. https://doi.org/10.1037/10011-000

Hagstofa Íslands. (2021). Innflytjendur 15,5% íbúa landsins (immigrants are 15.5% of the country's inhabitants) [Retrieved: 2022-08-28]. https://hagstofa.is/utgafur/frettasafn/mannfjoldi/mannfjoldi-eftir-bakgrunni-1-januar-2021/

Helgadóttir, S., Svavarsdóttir, Á., Rögnvaldsson, E., Bjarnadóttir, K., & Loftsson, H. (2012). The Tagged Icelandic Corpus (MÍM). *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, 67–72.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLOS ONE*, *10*(7), 1–23. https://doi.org/10.1371/journal.pone.0120644

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. *International Review of Applied Linguistics in Language Teaching*, *54*(3), 257–289. https://doi.org/doi:10.1515/iral-2015-0022

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P., & Hunston, S. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Nikulásdóttir, A. B. (2020). Models for automatic g2p for Icelandic (20.10) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/84

Nikulásdóttir, A. B. (2021). Icelandic pronunciation dictionary for language technology 21.10 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/154

Rögnvaldsson, E. (2015). Pronunciation dictionary for Icelandic [Retrieved: 2022-08-22]. http://www.malfong.is/index.php?lang=en&pg=framburdu

Rögnvaldsson, E. (2020). Icelandic pronunciation [CLARIN-IS]. http://hdl.handle.net/20.500.12537/82

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How Effective Are Intentional Vocabulary-Learning Activities? A Meta-Analysis. *The Modern Language Journal*, *104*, 715–738.

Xu, X., & Ingason, A. K. (2021). Developing Flashcards for learning Icelandic. *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, 55–61. https://aclanthology.org/2021.nlp4call-1.5.

Xu, X., Ingason, A. K., Kolka, V. T., Kovalova, A., & Kristínardóttir, I. (2023). Multilingual flashcards with 4,000 most common Icelandic words (IceFlash4K) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/308

Developing Manually Annotated Corpora for Teaching and Learning Purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene (the CrowLL Project)

Tanara Zingano Kuhn
University of Coimbra, Portugal
tanarazingano@outlook.com

Carole Tiberius
Dutch Language Institute, Netherlands
carole.Tiberius@ivdnt.org

Špela Arhar-Holdt
University of Ljubljana, Slovenia
Spela.Arhar-Holdt@ff.uni-lj.si

Kristina Koppel
Institute of the Estonian Language, Estonia
kristina.koppel@eki.ee

Iztok Kosem
University of Ljubljana/ Jožef Stefan Institute, Slovenia
iztok.kosem@ff.uni-lj.si

Rina Zviel Girshin
Ruppin Academic Center, Israel
rinazg@ruppin.ac.il

Ana R. Luís
University of Coimbra, Portugal
aluis@fl.uc.pt

Abstract

The CrowLL project seeks to provide manually annotated corpora for teaching and learning purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene, as a contribution to the Manually Annotated Corpora Family of resources available in CLARIN. Corpus sentences are annotated as "problematic" or "non-problematic" from the point of usage for pedagogical purposes. Sentences labelled as problematic also have annotations defining the category of the problem (offensive, vulgar, sensitive content, grammar/spelling problems, incomprehensible/lack of context). For each language, the corpus consists of 10,000 sentences, annotated by language experts. Additionally, we have developed a gamified solution for further corpus growth by using these annotated corpora as "seed corpora" to start the crowdsource-supported development of larger corpora. Annotation guidelines and game code will be published after the end of the project, thus enabling the application to other languages. The manually annotated corpora will allow language teachers, materials developers and lexicographers to use the labels to (de)select content/structure that is considered inappropriate or not (yet) suitable for the type of language learners involved. In addition to pedagogical goals, these corpora can also be used within NLP as datasets to train machine learning algorithms.

1 Introduction

This paper reports on the project Manually Annotated Corpora for Teaching and Learning Purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene (the CrowLL Project), which started on 01 September 2022 and finished on 31 August 2023. The project received funding from the CLARIN Resource Families Project (CRF).

The CrowLL project was born under the umbrella of the European Network for Combining Language Learning with Crowdsourcing Techniques (enetCollect) COST action[1] (CA 16105, 2017–2021) where a group of researchers worked towards building pedagogical corpora for teaching and learning purposes of Brazilian Portuguese, Dutch, Estonian, and Slovene. Corpora can be used to develop authentic language learning materials but might include sensitive content or offensive language, in addition to exhibiting structural (grammar, spelling) problems. Although such occurrences are unquestionably authentic, it is recommended that corpus examples are carefully monitored before applied in educational settings to flag inappropriateness, thus leaving the choice of use of certain examples to the needs and context of the use of teachers and didactic material developers. Monitoring corpus contents manually is extremely time-consuming and expensive, automating the process also has limitations. Hence a solution must be found to streamline human verification of examples.

In order to achieve our goal, the CrowLL project was initiated. The project aims to contribute to the Manually Annotated Corpora Family available in CLARIN by providing manually annotated corpora of Brazilian Portuguese, Dutch, Estonian, and Slovene – each containing 10,000 sentences annotated by members of our research group who are all language experts. Sentences in the corpora are marked with Y if the sentence was considered to be "problematic" for teaching the language and N if considered to be "non-problematic". All the problematic sentences additionally have labels indicating the category of the problem (offensive, vulgar, sensitive content, grammar/spelling problems, incomprehensible/lack of context). In addition to the manually annotated corpora, we have also developed a gamified solution for further corpus growth by using these annotated corpora as "seed corpora" to start the crowdsource-supported development of larger corpora.

The aims of this paper are to present the structure of the project and report some of the challenges that were faced in its development. In section 2, we describe how we organised our tasks in Work Packages, while in section 3 we discuss some of the decisions that had to be made and reflect on a few limitations. Finally, we conclude by showing how our project aligns with CLARIN's strategies and how it can contribute to several areas of knowledge.

2 Work Packages

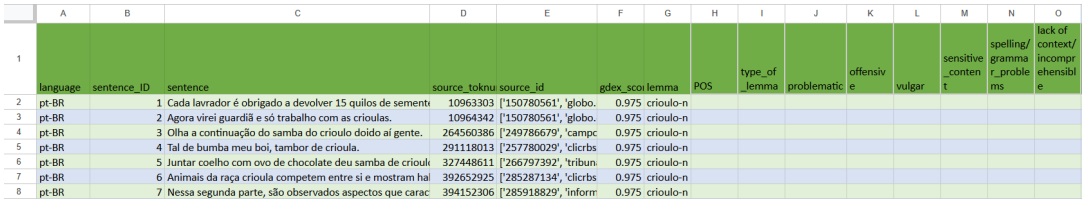
The project is organised in four Work Packages (WP). In this section, we describe the tasks carried out in each one of them.

In *WP 1: Preparing data and guidelines for corpus annotation*, we defined the criteria for the content of the seed corpora, extracted and prepared sentences for all the participating languages and created the annotation guidelines in all the languages (Brazilian Portuguese, Dutch, Estonian, and Slovene), including a translated version to English to allow scalability to other languages.

The annotated data consists of sentences that were automatically extracted from selected source corpora, primarily reference or web corpora for the participating languages. In our previous work (see Zingano Kuhn et al., 2021), we already identified available corpora from which the datasets for our work could be extracted and established that the extraction method would involve the use of specially-devised GDEX configurations (Kilgarriff et al., 2008; Kosem et al., 2019), blacklists and a lemma list that is common to all the participating languages (to allow comparability of the results). In the current project, we further specified these criteria (see Zingano Kuhn et al., 2022 for a detailed description) and extracted 10,000 sentences from each of the source corpora.

The data was prepared for two levels of annotation: 1. problematic/non-problematic from the point of usage for pedagogical purposes; 2. for each problematic sentence, the category of problem: offensive, vulgar, sensitive content, grammar/spelling problems, and incomprehensible/lack of context (Figure 1).

[1] https://www.cost.eu/actions/CA16105/



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	language	sentence_ID	sentence	source_toknu	source_id	gdex_sco	lemma	POS	type_of_lemma	problematic	offensive	vulgar	sensitive_content	spelling/grammar_problems	lack of context/incomprehensible
2	pt-BR	1	Cada lavrador é obrigado a devolver 15 quilos de sement	10963303	['150780561', 'globo.	0.975	crioulo-n								
3	pt-BR	2	Agora virei guardiã e só trabalho com as crioulas.	10964342	['150780561', 'globo.	0.975	crioulo-n								
4	pt-BR	3	Olha a continuação do samba do crioulo doido aí gente.	264560386	['249786679', 'campc	0.975	crioulo-n								
5	pt-BR	4	Tal de bumba meu boi, tambor de crioula.	291118013	['257780029', 'clicrbs	0.975	crioulo-n								
6	pt-BR	5	Juntar coelho com ovo de chocolate deu samba de crioul	327448611	['266797392', 'tribun	0.975	crioulo-n								
7	pt-BR	6	Animais da raça crioula competem entre si e mostram hal	392652925	['285287134', 'clicrbs	0.975	crioulo-n								
8	pt-BR	7	Nessa segunda parte, são observados aspectos que carac	394152306	['285918829', 'inform	0.975	crioulo-n								

Figure 1. Annotation spreadsheet for Portuguese.

Annotation guidelines with a description of these categories, illustrated by corpus examples, and an explanation of the metadata are published, together with the annotated corpora, as open data in PORTULAN CLARIN[2].

For further increase of these corpora, we propose a gamified crowdsourcing approach (Zviel-Girshin et al., 2021) where the community annotates the potentially problematic corpus sentences by playing a simple matching game, which will be demoed at the conference. This game will be first available as a single-player mode: players will get scores when their answers match previous answers given by other players. This was part of *WP2: Developing a gamified solution for further corpus growth.* By streamlining annotation and including more participants in the process, larger amounts of data can be manually processed. Under the umbrella of the enetCollect COST action (CA 16105), a prototype of the crowdsourcing game had already been developed. In the CrowLL project, the prototype was analysed, and improvements were implemented. At the time of writing, an initial dataset with 100 sentences per language has been imported into the game, and preliminary tests are taking place. These tests involve analysis of the annotation process (WP3), gameplay aspects (scoring mechanisms, incentive strategies, player experience), and interface structure. When the game is officially launched, the corpora with 10,000 manually annotated sentences (WP3) will have been imported to the game to be used as "seed corpora", i.e., so that players' answers can be matched to existing answers (experts' annotations). With this game, the definition of whether a sentence is problematic or not, to which category of problem it belongs, and what constituent part of the sentence is problematic will emerge from the majority consensus.

The code for gamified annotation, which will be published on Github as open access under Apache 2.0 licence, will be included among the project results. In the future, researchers wanting to create such annotated corpora for their language can choose either the expert approach (the annotation guidelines), or/and opt for crowdsourcing (the game).

In *WP3: Creation of the annotated corpora*, we annotated the corpora, finalised the guidelines and prepared the results in Excel format. In this Work Package, each corpus (one per language) of 10,000 sentences extracted from the source corpora (WP1) was manually annotated, following the annotation guidelines. Each sentence was first annotated as problematic or non-problematic from the point of usage for pedagogical purposes. Problematic sentences were further annotated with one or more categories labelling the nature of the problem. The annotation process was conducted by language experts. The final step of the process involved depositing the resources into PORTULAN CLARIN.

The last Work Package, *WP4: Documenting and modelling the results for further use*, consists of modelling the methodology for further use and Dissemination (blog post on the CLARIN webpage, CLARIN Café, project webpage, social media, etc).

After the project, we will document and describe all parts of the methodology in the form of a conference paper and, in succinct version, as a blog post on the CLARIN webpage; in this document, we will include a plan for further growth of our corpora with the help of the crowd, as well as lessons learned for researchers who are interested in repeating the process in their own languages. The results of the project will be communicated and disseminated to different audiences and with different objectives. For the specialised audience, we plan to have a CLARIN Café and prepare documentation, which will enable and stimulate the extension of the new resource family to other languages. For the general public, we will publish the results of the project, carefully expressed in a non-specialised manner, on the website of our project[3] to promote the connection between academia and society.

[2] https://hdl.handle.net/21.11129/0000-0010-05DA-3
[3] https://ucpages.uc.pt/celga-iltec/crowll/

3 Discussion

One of the crucial guidelines for choosing our source corpora was that they were at least in some part openly available. While for Estonian and Slovene we could use corpora that have been carefully compiled in the context of other projects, with rich metadata and advanced annotation, Dutch and Portuguese had to resort to the automatically-compiled Timestamped JSI web corpora (a family of web corpora created by the Jožef Stefan Institute in Slovenia from IJS newsfeed for 18 languages (Trampuš & Novak, 2012) with no human curation). These corpora have been uploaded and POS-tagged by the Sketch Engine team and we used the version of these corpora in Sketch Engine to extract candidate example sentences for Portuguese and Dutch. It should be acknowledged that these differences in the development of the resources might influence the quality of the input data (extracted sentences), with consequent reflection on the quality of the output data (annotated sentences).

Another potential issue concerns linking the annotated sentences with the source data. For this, we kept the metadata which is associated with these sentences in the Sketch Engine. Although this does provide us with an identifier, it is unfortunately not a persistent identifier as the link may break if at some point in the future, the corpora will be reindexed or even deleted in Sketch Engine. For the Estonian and Slovenian data, this is not an issue as we use (the in-house version of) Gigafida 2.0 (Krek et al., 2020) for Slovene, and the Estonian National Corpus 2021 (Koppel & Kallas, 2022) for Estonian.

Finally, it should be noted that the boundaries between some of the proposed problem categories are not always easy to draw. Different decisions are possible even when sentences contain words that are clearly (on the surface and in the vast majority of the meanings) offensive or vulgar, for example: *nigger*, *whore*, *bitch*, *retarded*, *to fuck*, *to piss*. These words would typically make it to blacklists, and blacklist-based methodology would automatically filter out the sentences before they would be included in any teaching material. Here, we included them to test the hypothesis of whether all extracted sentences would be marked as inappropriate by the annotator(s) and the crowd. Sentences containing such 'blacklisted' words should clearly not appear in learning material without a label as they may cause discomfort or upset or embarrass people. However, if in a specific context a word such as *nigger* pops up and a teacher would like to explain the word and the connotations associated with it by means of example sentences, a sentence such as *Het Rijksmuseum heeft termen als "neger" verwijderd*. ('The Rijksmuseum has removed terms such as *nigger*) could (in our opinion) be used. These sentences are now annotated as problematic and labelled as sensitive. Sentences containing these words and that are clearly not suitable in any context are labelled as offensive. We believe the use of crowdsourcing for the annotation of examples like this will provide useful insights on the wider perception of the selected problem categories.

4 Conclusion

The CrowLL project is very much aligned with CLARIN's strategies of providing data for facilitating research in data science and artificial intelligence, as well as social sciences and humanities. The project is also strongly committed to the FAIR principles, promoting open science and scalability of planned results. One of the strengths of the project is innovation, both in terms of data and methodology. Moreover, the corpus resources produced by the current project will extend the scope of the CLARIN Resource Families initiative, more particularly, the section on manually annotated corpora. Creating annotated corpora of several languages will enable research on linguistic levels as well as training of tools, while accompanying documentation, gamification and presentations will facilitate further expansion of this type of data collection to other languages. The resulting corpora will also support language teachers by providing labelled example sentences that can be used in class.

Acknowledgements

This study has been supported by the CLARIN Resource Families Project Funding. The authors acknowledge the financial support received from the Portuguese national funding agency, FCT – Foundation for Science and Technology, I.P. (grant number UIDP/04887/2020) and the Slovenian Research Agency (research core funding No. P6-0411, Language Resources and Technologies for Slovene, and project funding No. J7-3159, Empirical foundations for digitally- supported development of writing

skills). The research received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 731015.

References

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX international congress, Vol. 1*. 425–432. https://tinyurl.com/yckr9w8s

Koppel, K., & Kallas, J. (2022). Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics*, 18, 207−228. https://doi.org/10.5128/ERYa18.12

Kosem, I., Koppel, K., Kuhn, T. Z., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography, 32*(2), 119−137. https://doi.org/10.1093/ijl/ecy014

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A.Ž, Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ur.): *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Palais du Pharo, Marseille, France*. ELRA - European Language Resources Association, 3340−3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Zingano Kuhn, T., Arhar Holdt, Špela, Kosem, I., Tiberius, C., Koppel, K., & Zviel-Girshin, R. (2022). Data preparation in crowdsourcing for pedagogical purposes: The case of the CrowLL game. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research*, *10*(2), 62–100. https://doi.org/10.4312/slo2.0.2022.2.62-100

Zingano Kuhn, T., Todorovic, B. Š., Arhar Holdt, Š., Zviel-Girshin, R., Koppel, K., Luís, A. R., & Kosem, I. (2021). Crowdsourcing pedagogical corpora for lexicographical purposes. *Proceedings of EURALEX 2020 Conference, Volume II*, 771 – 779. http://https://euralex2020.gr/proceedings-volume-2/

Zviel-Girshin, R., Kuhn, T. Z., Luís, A. R., Koppel, K., Šandrih, B., Arhar Holdt, Š., Tiberius, C., & Kosem, I. (2021). Developing pedagogically appropriate language corpora through crowdsourcing and gamification. In: Zoghlami, N; Brudermann, C.; Sarré, C.; Grosbois, M.; Bradley, L.; Thouësny, S. *CALL and professionalisation: short papers from EUROCALL 2021*. Research-publishing.net. https://doi.org/10.14705/rpnet.2021.54.9782490057979

Trampuš, M., & Novak, B. (2012). Internals of an aggregated web news feed. *Proceedings of 15th Multiconference on Information Society*, 221–224.