CLARIN Annual Conference Proceedings

2024

Edited by

Vincent Vandeghinste and Thalassia Kontino

15 – 17 October 2024
Barcelona, Spain

Please cite as:
CLARIN Annual Conference Proceedings, 2024. ISSN 2773-2177 (online).
Eds. Vincent Vandeghinste and Thalassia Kontino.
Barcelona, Spain, 2024.

Programme Committee

Chair:

- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (NL) & KU Leuven (BE)

PC Subcommittee:

- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Cristina Grisot, University of Zurich (CH)
- 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)
- Krister Lindén, University of Helsinki (FI)
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)
- Tanja Wissik, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Gijsbert Rutten, Leiden University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- German Rigau, HiTZ, the Basque Center for Language Technology (ES)
- 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 (NL) & 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
- Tomaž Erjavec, SI
- Cristina Grisot, CH
- Eva Hajičová, CZ
- Krister Lindén, FI
- Monica Monachini, IT
- Tanja Wissik, 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
- Martin Critelli, IT
- Riccardo Del Gratta, IT
- Maria Gavriilidou, GR
- Enikő Héja, HU
- Jesse Holmes, EE
- Penny Labropoulou, GR
- Rooweither Mabuya, ZA
- Katja Meden, SI
- Pärtel Lippus, EE
- Deon du Plessis, ZA
- Tõnis Nurk, EE
- Valeria Quochi,
- Michael Rießler, EE
- Mateja Jemec Tomazin, SI
- Bram Vanroy, BE
- Benito Trollip, ZA

CLARIN 2024 submissions, review process and acceptance

- Call for abstracts: 23 January 2024 first call published on CLARIN website, disseminated, and submission system open
- Submission deadline: 26 April 2024
- In total 60 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 10 June 2024
- Notifications to authors: 17 June 2024
- 42 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2024/clarin-annual-conference-2024.

Preface

It is with great pleasure that we present the Proceedings of the CLARIN Annual Conference 2024. The conference is held in Barcelona, from 15 till 17 October 2024.

CLARIN conferences have been organised since 2012 and are the place where the people behind the pan-European research infrastructure that CLARIN is, meet with each other, with users of the CLARIN infrastructure and with tool and resource builders that provide their tools to the infrastructure. Together we form the CLARIN community. At this year's conference we welcome 194 registered on-site participants and another 114 participants online.

We have thematically organised the 20 oral presentations (and the written proceedings) in 5 themes: Resources and usage, Education, Core CLARIN Infrastructure, Metadata, and Centres and Resource Families. Additionally we have 22 poster presentations, totalling 42 peer reviewed papers.

We would like to acknowledge the work done by the PC Committee, by the National Coordinators and by the other reviewers to ensure the quality of the papers.

A new thing at this year's conference is the CLARIN 101 Workshop, providing some background information for those people that are new to CLARIN.

We are also very pleased with both keynote speakers. Maite Melero will talk about *The Future of Language (and Cultural) Diversity in the Age of AI* and Steven Bird will give a talk titled *Making it Meaningful*. We hope that both these talks will provide us with inspiration and reflection in an era where language technology and artificial intelligence developments have become omnipresent.

We hope you enjoy your stay in Barcelona, that you have an inspiring conference and that you may have interesting interactions with other members of the CLARIN community.

Vincent Vandeghinste
Programme Committee Chair
National Coordinators Forum

Thalassia Kontino
CLARIN ERIC

Table of Contents

Resources and usage

Education

Core CLARIN Infrastructure

Metadata

Centres and Resource Families

Posters

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

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

Abstract

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

1 Introduction

It is not new to use the term *infrastructure* for the organisation of terminological collaboration and terminological activities (Pilke et al. 2021, 101). Already Galinski (1998) described an infrastructural approach to terminology, dividing terminological infrastructures into horizontal and vertical infrastructures. The horizontal infrastructure includes five elements: “terminology (planning) policy, terminology creation centres, terminology information and documentation centres, terminology associations and corporate cooperation groups led by the private sector” (Galinski, 1998). The vertical infrastructure concerned the different ways of caring out terminological activities within different domains (Galinski, 1998). An adapted version of this infrastructural model (Galinski & Giraldo, 2023) is used in a cooperation project between Austria and Mongolia[1]. In these existing infrastructural models for terminology, research infrastructures are not explicitly included, however there are several connecting points between terminology work and research infrastructures (e.g., Andersen & Gammeltoft 2022, Wissik & Declerck 2020, Wissik, 2022). Stakeholders in terminology work can be on the one hand data providers and on the other hand they can be users of data, tools and services provided by RIs and benefit from the knowledge sharing infrastructure to exchange knowledge and promote collaboration.

However, there is little insight into the role and use of such research infrastructures, in particular CLARIN and DARIAH, within the community of stakeholders involved in terminology work, especially in institutional settings, besides some case studies (e.g., Andersen & Gammeltoft, 2022). So, this paper wants to close this gab and explores the role of research infrastructures with regard to terminology work in institutional settings (academic and non-academic) based on qualitative interview data. The paper is structured as follows: after an introduction, the research method is described, and the results are discussed. The contribution focuses on resources (e.g., corpora) and repositories as possible links between research infrastructure and the community of stakeholders involved in terminology work, mentioning also other possible areas of interaction such as training materials and tools.

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

2 Method

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

To gain insight in this area, 15 semi-structured expert interviews were conducted with individuals involved in terminology workflows in different institutional settings in different roles to better understand terminology workflows in the digital age (Wissik, forthcoming a). Since the study focuses on terminology practices in institutional settings, it was important to interview individuals from different types of institutions while covering most common types of terminology work: The interview participants came from academic and non-academic institutions. There were both institutions from member countries and non-member countries (see Table 1).

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

Interview	Role	Type of institution	CLARIN Member	DARIAH Member
INT 1	Developer / IT Expert	Academic/ Research	NO	YES
INT 2	Technology Manager	Academic / Research	NO	YES
INT 3	Head of Terminology Unit / Terminologist	Academic / Research	YES	YES
INT 4	Member of Terminology Committee	Academic / Research	NO	YES
INT 5	Head of Terminology and Legal Translation Unit, Deputy Director for Development	Administration (state level)	YES	Cooperating Partner
INT 6	Terminologist	Administration (regional level)	YES	YES
INT 7	Terminologist	Academic / Research	YES (At the time of the interview not yet full member, only K-Centre)	YES (At the time of the interview not yet full member)
INT 8	Head of project management unit / Terminologist	Administration (regional level)	YES (At the time of the interview not yet full member, only K-Centre)	YES (At the time of the interview not yet full member, only K-Centre)
INT 9	Terminologist	Academic / Research	YES (At the time of the interview not yet full member, only K-Centre)	YES (At the time of the interview not yet full member, only K-Centre)
INT 10	Terminology Coordinator / Terminology Manager	European Institution	Not applicable	Not applicable
INT 11	Head of Terminology Unit / Terminologist	International Organization	Not applicable	Not applicable
INT 12	Technology Manager	European Institution	Not applicable	Not applicable
INT 13	Head of Terminology Unit / Terminologist	International Organization	Not applicable	Not applicable
INT 14	Terminology Coordinator / Terminology Manager	Academic / Research	YES	Cooperating Partner
INT 15	Head of Terminology Unit / Terminologist	Administration (regional level)	OBSERVER	YES

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

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

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

3 Results

3.1 Corpora

Using corpora in terminology work is not new (Bowker, 1996). When doing ad hoc terminology work, for example, answering requests from query services (Žagar Karer & Fajfar 2023), terminologists usually resort to already existing corpora:

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

When doing systematic terminology work for a specific domain in order to create for example, a specialized dictionary or enrich a terminology database with a new domain, terminologists also create specialized corpora from scratch: “In the beginning when we started compiling […] dictionaries we always prepare a specialized corpus of the texts that the experts give them [the texts] to us and then we the terminologist, prepare wordlist” (INT 3). The snippets from the interview show, that in institutional terminology work on the one hand already existing corpora are consulted and on the other hand specialized corpora are created from scratch. However, most interview participants create the corpora for internal use only. They do not publish them or deposit them in a repository: “We usually keep it as a working material. It's more like really like a stage in preparing a dictionary, it is not annotated with POS [part of speech]. I would take us too much time for this.” (INT3). Sometimes they are also shared outside the organization but they are not published:

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

3.2 Repositories

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

For this contribution we only analysed the use of data repositories that are related to RIs. Only one interview participant reported, that their data is stored in a national CLARIN repository (INT 3) and one interview participant reported, that they had recently talked with a national CLARIN representative also about the possibility of archiving the data in a CLARIN repository in the future (INT 14).

3.3 Collaboration with Research Infrastructures such as CLARIN or DARIAH

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

4 Discussion and Conclusion

Regarding the use of corpora, most of the interviewees reported, that they were using already existing corpora. In these cases, also corpora from CLARIN national consortia were mentioned and used. However, not all interviewees were aware of the available resources through CLARIN and DARIAH. Regarding the creation of their own specialized corpora, nearly all of the interview partners mentioned, that they are not publishing these data sets, because for example they are not annotated with part of speech. So, RIs could provide information, that also not annotated corpora are a valuable resource for sharing and reusing.

Regarding the publication of terminological data in data repositories as an additional channel was not a very common practice at the time of the interviews. In fact, only one participant mentioned, that they deposit their data, where copyright allows it, in a national CLARIN repository. So, in this respect, awareness raising regarding the use of data repositories in this community would be needed, in both academic and non-academic settings.

Most participants from non-academic institutions did not have collaborations with RIs and some of them were not familiar with CLARIN and DARIAH, even though their institutions were located in a member country or the country had at least a K-Centre (in case of CLARIN). These results are not so surprising, as the priority within the RIs so far was to reach out to academic users and to broaden the academic user base. However, recently RIs started to engage with non-academic communities as well, such as the public sector. In the case of CLARIN, with the help of dedicated K-Centres, training materials could be created, that specifically target the terminology community. Furthermore, an ambassador from the terminology community within the public sector could be used to get engage with others in the public sector. By recruiting ambassadors also from the public sector, the already existing and successful CLARIN ambassadors programme could be used to reach out to the public sector. Furthermore, CLARIN and DARIAH could engage with this community via terminology or language associations, where they are members. Furthermore, terminology communities that already benefit from RIs like in Norway (Andersen & Gammeltoft, 2022) could be used to showcase the benefits of RIs in terminology work. Another finding of the analysis was, that interviewees that were aware of CLARIN and/or DARIAH, reported, that the specific academic sub-unit involved in terminology work was not having links with these RIs. It is clear, that institutions such as universities are complex but it could be worthwhile to investigate, how to best target relevant academic users in sub-units of institutions that are already cooperating with RIs. Furthermore, several tools to manage, edit and visualize data play an important role in terminology work (e.g., terminology management systems, corpus management tools, term extraction tools) which could be also a possible area of interaction with research infrastructures.

To sum up, the case study has shown, that the role of RIs in terminology work in institutional settings has potential but is still expandable. Furthermore, the contribution has discussed the possibility to expand the CLARIN ambassador programme and recruit also ambassadors from the public sector. So, stakeholders in terminology work for example in public administration could act as ambassadors to engage with non-academic communities who could benefit from the data, services and knowledge provided by RIs.

Acknowledgement

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

References

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

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

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

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

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

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

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

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

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

Wissik, T. (forthcoming a). Dimensions of sustainability in terminology practices in institutional settings. *Terminology Science & Research*, 27.

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

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

Using the Icelandic Gigaword Corpus to Explain Lifespan Change

Lilja Björk Stefánsdóttir
University of Iceland
lbs@hi.is

Anton Karl Ingason
University of Iceland
antoni@hi.is

Abstract

In this paper, we demonstrate research on syntactic lifespan change in the speech of an Icelandic MP, using data from the Icelandic Gigaword Corpus. Our study exemplifies how advances in language technology infrastructure can play an important role in linguistic studies, providing comprehensive linguistic data that would have been impossible to acquire only a few years ago.

1 Introduction

In this paper, we describe a case study where we use the Icelandic Gigaword Corpus (Steingrímsson et al., 2018) in order to examine syntactic lifespan change in the speech of an Icelandic MP, with an emphasis on the period before, during, and after the Icelandic economic crash of 2008. Our study exemplifies the benefits of using a CLARIN language resource for big data humanities.

We examine the variable use of the syntactic process of Stylistic Fronting (SF), a stylistic indicator associated with formal speech in Icelandic, throughout the career of an Icelandic MP, Bjarni Benediktsson. We observe Benediktsson having a high rate of SF usage during his first years as an MP, a pattern that is disrupted in the year 2008, where the rate suddenly drops following the Icelandic economic crash of 2008. Benediktsson's rate of SF continues to drop in the years following the economic crash, an interesting development as it occurs at the same time as his party, The Independence Party, is in a very weak position, having lost their seat in government and facing the lowest support in the history of the party. We attribute this decline to a dramatic change in Benediktsson's Linguistic Market Value (LMV) in the sense of D. Sankoff and Laberge (1978); greater contextual importance of language correlates positively with the use of more formal speech variants. This temporary change is then reversed in 2012 when the rate of SF increases again, during a time where the aftermath of the economic crash is mostly over and Benediktsson's status, as well as his party's status within the parliament is getting stronger.

The findings from our study demonstrate how a fine-grained view of syntactic lifespan change yields insights about status-associated usage as interrelated aspects of the social dimension of language. Our findings also provide evidence of the importance of a high-definition approach (Stefánsdóttir & Ingason, 2018, 2019), that is, using comprehensive and continuous linguistic data derived from corpora, due to the complex and fluctuating nature of individual lifespan change.

2 Background

In the past two decades, there has been considerable growth in research on changes in how individuals speak over the course of their lives, generally referred to as lifespan change. Most studies on lifespan change are not high-definition studies, meaning they are commonly based on comparing two periods in the individuals's life. In some cases, samples from several periods are combined into one and then compared to another perio, like in MacKenzie's (2017) study on David Attenborough's speech. Sankoff's (2004) study on phonological variation among members of the film series "7 Up" is unusually fine-grained because the same individual was examined five times at seven-year intervals, therefore relying on a higher time resolution than most studies on individual lifespan change. Yet, we cannot see how the changes develop accurately because seven years passed between points in time when the same individual was observed. The study by Arnaud (1998) is similar to the present one as it uses a corpus as a source for tracking syntactic change. Although the study certainly does have a higher time resolution than many comparable studies, it nevertheless leaves room for enhanced detail, for example, the study is exclusively based on written texts, and its quantitative findings are not based on a well-defined envelope of variation but rather a more coarse density measure. Furthermore, the data are grouped into 5-year periods, yielding a maximum of 11 readings per speaker. Therefore, we believe that there is much to be learned from our present study, which looks at a **continuous year-by-year account** of well-understood linguistic variable that spans several years and is derived from spoken language transcriptions, thus having a very high time resolution and a robust connection with spoken Icelandic.

Two of our previous studies (Stefánsdóttir and Ingason, 2018, 2024) on Icelandic MPs during the financial crash are examples of the benefits of this kind of methodology. The former Icelandic Minister of Finance, Steingrímur Sigfússon, showed an upward shift in his use of SF during the economic crash, which we attributed to a dramatic change in his Linguistic Market Value (LMV), as he was in a position of great responsibility as a Minister of finance during an economic crisis. Another MP, Þorgerður Gunnarsdóttir, showed a different reaction to the crisis as the rate of SF dropped during the period, especially between the years 2008–2009, when the rate went from 78% to 50%. This change is interesting because Gunnarsdóttir had been a minister in the so-called crash-government, that is, the government that had been serving during and in the years before the economic crash, which eventually collapsed in early 2009 after the biggest public protests in Icelandic political history. We attributed this drop to a change in her LMV, as she had gone from being a minister in the government to an MP in the opposition as well as being a member of a party that suddenly had become very unpopular.

All of our studies on Icelandic MPs are high-definition studies, meaning that they rely on a very high time resolution, with the number of readings matching the number of years the MPs have been in office. For Sigfússon and Gunnarsdóttir, we had a total of 38 and 20 readings, respectively, giving us a clear overview of changes and their development with no gaps between readings.

2.1 The Stylistic Fronting variable

Stylistic Fronting is an optional movement process, found in Icelandic, of a word or a phrase into a subject gap (Angantýsson, 2017; Maling, 1980; Thráinsson, 2007; Wood, 2011).

(1)	*Bækur*	*[$_{CP}$ sem*	*{eru*	*lesnar*	*(No SF)*	*//*	*lesnar*	*eru (SF)}*	*til*	*skemmtunar]*	*eru*	*bestar.*
	books	[$_{CP}$ that	are	read		//	read	are		for	entertainment]	are	best

‘Books that are read for entertainment are the best ones.’

The contrast between the word orders with and without SF illustrates the optionality. The relative clause has a subject gap and thus SF can apply and move the non-finite main verb in front of the finite auxiliary. SF has no effect on truth-conditional meaning, and its only clear meaning component is a sociolinguistic one; the movement is associated with formal style. SF is found in both main clauses and subordinate clauses, as long as the subject is not phonologically overt. The phonological subject gap condition holds for relative clauses with extracted subjects, as above, and impersonal main clauses where there is no overt subject. Although full phrases can be stylistically fronted, we only focus on the cononical case here, as we limit the scope of the study to word orders involving the complementizer *sem* that introduces Icelandic relative clauses (e.g., by excluding frontable elements other than non-finite main verbs) and finite auxiliaries and non-finite main verbs in either of the two possible orders.

We do this to control for factors that can condition the use of SF, building on findings from Wood (2011) who showed that prosodic factors and syntactic category can affect the rate of SF. Obviously, this does not include all cases of SF but allows us to extract a well-defined envelope of variation with high accuracy (where SF application and non-application are accounted for).

3 Detecting patterns in the Icelandic Gigaword Corpus

The Icelandic Gigaword Corpus consists of 2429 million running words of text, and a part of the corpus is parliament speeches. This resource allows for a high-definition time resolution; thus, we can observe gradual changes over time where the time axis is continuous. We wrote a Python script that analyzed a part of the corpus, the parliament speeches by Benediktsson between 2003–2021, and we extracted sequences with a relative complementizer followed by a finite verb and a non-finite one in either of the two possible word orders. The corpus includes audio files and a transcription, making observations for accuracy possible as we can listen to each audio file and verify the transcription. The patterns that we search for are very reliable as confirmed by our manual checks. As mentioned, we only collected subject relatives with a potential for SF of a non-finite main verb. This provided us with 2729 tokens of the SF variable, with each token coded for SF application and the year of the speech.

4 Lifespan change in Benediktsson's speeches

Figure 1 shows Benediktsson's use of SF across his career. In the early years, Benediktsson's rate of SF is relevatively high, with the average of 91.16% in the years 2003–2007. This pattern is not an unexpected one since Benediktsson was a new MP at the time and the situational effect of his surroundings and new status is likely to have caused him to be become more aware of his language use, which positively correlates with a frequent use of formal variants such as Stylistic Fronting, according to Labov's (1972) attention-paid-to-speech model. In addition to situational effects, the high rate of SF during this period can also be interpreted in terms of the so-called Linguistic Marketplace (D. Sankoff and Laberge, 1978), a long term style predictor that explains the relationship between a speakers linguistic behavior and their Linguistic Market Value (LMV). Individual's LMV is highly connected to their social status, which is shaped by personal and professional experience over the years, and a high LMV correlates positively with the probability of using formal variants. Benediktsson's LMV was high during this period, both as an MP but also as an MP for a party that was in government and therefore, in a strong position within the parliament, resulting in his speech becoming formal.

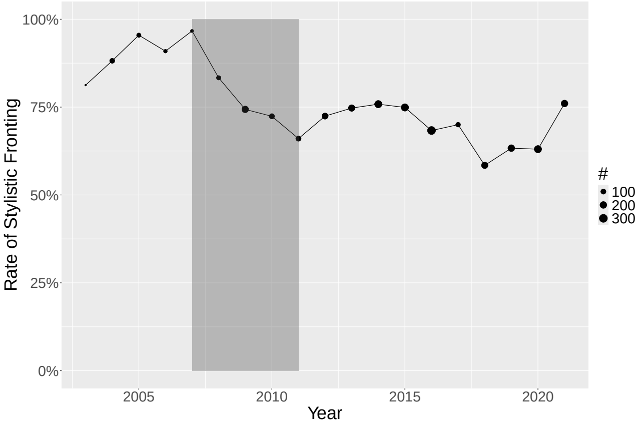


Figure 1: Evolution of SF in Benediktsson's career.

However, this trend is reversed in 2008 when the rate of SF suddenly drops, going from 96.7% in 2007 to 83.3% in 2008. The rate continues to drop in the following years (2007–11 highlighted), reaching its lowest rate in 2011 at 66.0%. This decline in the use of SF is interesting because it occurs almost simultaneously with the Icelandic economic crash of 2008 and during the crash's aftermath which lasted till the year 2011. The economic crash took a big toll on Benediktsson's party, The Independence Party, as it had been a part of the so-called crash-government, which collapsed in early 2009, and lost a lot of support from its voters with many looking at the party as the culprit for the economic crash. We analyze this change in Benediktsson's linguistic behaviour as a reflection of a dramatic change in his LMV, due to his weak political position at the time.

This temporary change is again reversed in 2012, when the rate of SF increases again, when the aftermath of the financial crash is mostly over and Benediktsson, and his party, start to prepare for the upcoming elections, which took place a year later. The 2013 elections resulted in the Independence Party regaining their status as Iceland's biggest political party and Benediktsson took on the role of Minister of Finance in a right-wing government. Perhaps unsurprisingly, the rate of SF continues to be relatively high in the following years, reflecting Benediktsson's high LMV as a person in a position of power.

An analysis of Benediktsson's use of SF in the period before, during and after the financial crash in Iceland confirms that the various nuances of individual lifespan change can only be studied in a large digitized corpus. If we had, for example, only the first data point and the last, many crucial aspects of the development would have gone missing from the picture, no matter how carefully the data would have been collected. Note that analysis of lifespan change has to be evaluated on a case-by-case basis and, for example, in our study of Ásmundur Daðason (Stefánsdóttir & Ingason, Forthcoming), we believe that aspects of identity are more important than LMV. Also note that it is not trivial to decide how to evaluate statistical signficance in such a data set. Some pairwise differences between years are significant but we seek tools that evaluate differences in wiggly long-term trends. We leave a more accurate analysis for future work.

5 Summary

In this paper we used a CLARIN resource, a corpus containing the speeches of the Icelandic parliament, to analyze the formality levels of one politician, Benediktsson, over time. This builds on our previous work on other Icelandic Members of Parliament and continues to demonstrate the value of open access parliament data for the digital humanities. We found that fluctuations in the use of SF by Benediktsson reflects explanations given by the literature on style shift, as it connects with his attention-paid-to-speech and Linguistic Market Value over time.

References

Angantýsson, Á. (2017). Stylistic fronting and related constructions in the insular scandinavian languages. *Syntactic Variation in Insular Scandinavian*, *1*, 277.

Arnaud, R. (1998). The development of the progressive in 19th century English: A quantitative survey. *Language Variation and Change*, *10*(02), 123–152.

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

MacKenzie, L. (2017). Frequency effects over the lifespan: A case study of Attenborough's r's. *Linguistics Vanguard*.

Maling, J. (1980). Inversion in embedded clauses in Modern Icelandic. *Íslenskt mál*, *2*, 175–193.

Sankoff, D., & Laberge, S. (1978). The linguistic market and the statistical explanation of variability. In *Linguistic variation: Models and methods*. Academic Press.

Sankoff, G. (2004). Adolescents, young adults and the critical period: Two case studies from 'Seven up'. *Sociolinguistic variation: Critical reflections*, 121–139.

Stefánsdóttir, L. B., & Ingason, A. K. (2018). A high definition study of syntactic lifespan change. *U. Penn Working Papers in Linguistics*, *24*(1), 1–10.

Stefánsdóttir, L. B., & Ingason, A. K. (2019). Lifespan change and style shift in the Icelandic Gigaword Corpus. *Proceedings of CLARIN Annual Conference 2019*, 138–141.

Stefánsdóttir, L. B., & Ingason, A. K. (2024). *Lífsleiðarbreytingar á Alþingi* [The Annual Humanities Conference, University of Iceland. Reykjavík 8–9 October.].

Stefánsdóttir, L. B., & Ingason, A. K. (Forthcoming). Wiggly lifespan change in a crisis – contrasting reactive and proactive identity construction. *U. Penn Working Papers in Linguistics*.

Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., & Guðnason, J. (2018, May). Risamálheild: A very large Icelandic text corpus. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association (ELRA). https://aclanthology.org/L18-1690

Thráinsson, H. (2007). *The syntax of Icelandic*. Cambridge University Press.

Wood, J. (2011). Stylistic fronting in spoken Icelandic relatives. *Nordic Journal of Linguistics*, *34*(1), 29–60.

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

Steinþór Steingrímsson, Einar Freyr Sigurðsson, Björn Halldórsson
The Árni Magnússon Institute for Icelandic Studies
{steinst,einasig,bjornh}@hi.is

Abstract

Multiword expressions (MWEs) are generally problematic for machine-translation systems. In this paper, we (i) describe a set, available on CLARIN-IS, of appr. 1,000 idiomatic MWEs which have been translated into English; and (ii) evaluate – using both automatic and manual approaches – three MT systems' abilities to translate MWEs from Icelandic to English. We find that the MT systems evaluated commonly fail when translating idiomatic expressions.

1 Introduction

Multiword expressions (MWEs) are a frequent phenomenon in natural language and speech[1]. Proper handling of MWEs is important for various natural language processing (NLP) tasks, such as machine translation (MT), bilingual lexicon induction and information extraction. It is difficult to provide clear boundaries for what constitutes a MWE and what does not. The term can be used to describe fixed or semi-fixed phrases, compounds, idioms, phrasal verbs or collocations – in general, any sequence of words that acts as a single unit on some level (Calzolari et al., 2002).

In this paper, we introduce a set of approximately 1,000 Icelandic MWEs[2], along with their translations into English as well as structured information about their usage. We classify the MWEs in our dataset *idiomatic expressions*, i.e. idioms with an intended meaning that diverges from the literal meaning of the words constituting the expression, and therefore usually cannot be translated word for word. Machine-translation systems generally do not handle MWEs well, and even though they are an important part of generating fluent translations they can be a blind spot for traditional automatic evaluation approaches, such as BLEU (Papineni et al., 2002) or chrF++ (Popović, 2017). This applies especially in cases where there is more than one "right" answer, as the traditional lexical metrics cannot identify what goes wrong in a translation. The Icelandic MWE dataset was compiled for use with MT, and can be used either to augment training sets with sentence pairs containing common idiomatic expressions, or for evaluating the capabilities of MT systems to translate such expressions. We show how the dataset can be used to evaluate the capabilities of three machine translation (MT) systems to translate MWEs, by evaluating the systems in three different ways: using traditional automatic approaches, using automatic evaluation of MWE translations, and by manually evaluating the output.

2 Collecting the Multiword Expressions

The set of multiword expressions, distributed on the Icelandic CLARIN repository[3], contains approximately 1,000 Icelandic idioms processed from the ISLEX dictionary (Úlfarsdóttir, 2014). They are listed with their English idiomatic equivalent and literal meaning in both languages, as well as example sentences and keywords. The idioms are, in most cases, syntactically mobile, which is why case information is included.

The idioms were processed from a list of 4,000 MWEs in the ISLEX database. The idioms are ordered alphabetically according to the first keyword of each idiom and each line contains the following

[1] We thank three anonymous reviewers for valuable comments on the paper.

[2] http://hdl.handle.net/20.500.12537/275

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

categories: 1) Icelandic idiom; 2) English equivalent; 3) Meaning of the Icelandic idiom; 4) Meaning of the English idiom; 5) An example sentence with the Icelandic idiom; 6) An example sentence with the English idiom; 7) An example sentence with the meaning of the Icelandic idiom; 8) An example sentence with the meaning of the English idiom; 9) Keywords in the Icelandic idiom, lemmatized (in some cases in the plural). The first four categories contain type information, cf. for example, the idiom *rétta e-m hjálparhönd* 'lend someone a helping hand', which is listed as follows: <NP1-nom> rétta <NP2-dat> hjálparhönd; <NP1> lend <NP2> a helping hand; <NP1-nom> hjálpa <NP2-dat>; <NP1> help <NP2>.

Where more than one English equivalent, translation or sense are possible, alternatives are separated with a pipe symbol. Keep in mind that there is not always a 1-1 relation between the example sentences. For example, the Icelandic idiom *Það er fokið í flest skjól*, is translated as 'We're at the end of our tether', where the Icelandic expletive *það* 'it, there' makes way for a personal pronoun in English. The use of other symbols in the file is as follows: alternatives within the same segment are separated with a slash (/), e.g. the idiom *vera klár/tilbúinn í slaginn* ('be ready to rumble'), and optional parts of idioms are in parentheses, e.g. the idiom *bretta upp ermar(nar)* ('roll up one's sleeves') or *vera sjálfs sín(s) herra* ('be one's own boss').

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

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

For the MT evaluation, we process the data in a slightly different way than in the distribution file. We number each segment, and while we only use the example phrases and their translations, where there are alternatives within the segments we generate all possible pairs. The generated pairs then get the segment number and the evaluation results are weighted so that all segments in the dataset have the same weight in the final score. Furthermore, we add a list of words that should be included in the MT translation of the idiom, and that list is used for the automatic evaluation of idiom translation. The processed data, along with all scripts, are made available on GitHub[4].

3 Evaluating Machine-Translation Systems

When choosing which MT system to use for a given task, the ability to translate MWEs can be a deciding factor. It is therefore important to be able to test that ability. To this end, we run three evaluation experiments. First, we simply evaluate the MT output using traditional automatic approaches. We apply the common evaluation metrics BLEU and chrF++. Second, we devise a simple automatic approach that classifies translations in two groups: translations likely to have correctly handled the MWE and translations that failed to do so. Third, we manually evaluate all translations to be able to confirm or reject the adequacy of the automatic approach.

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

MT System	BLEU	ChrF	AutoIE (%)
Steingrímsson (filtered parallel data)	9.5	33.2	9.2
Miðeind (using backtranslations)	10.7	33.9	11.5
Google Translate	21.0	49.4	15.5

Table 1: Automatic evaluation of the three MT systems.

3.1 MT Systems

We compare three MT systems that translate from Icelandic into English. The first model is an Icelandic–English translation model (Símonarson et al., 2022) trained by the language-technology startup Miðeind, based on the mBART25 model (Liu et al., 2020). The Miðeind-model is trained on long-context texts, both authentic parallel texts and synthetic texts. The synthetic data comprise backtranslations from various sources, totalling over 150 million tokens. The second model is also based on mBART25. It is only trained on authentic parallel texts, and uses the set of training sentences compiled by Steinþór Steingrímsson using the most effective filtering approach reported in his thesis (Steingrímsson, 2023). The bulk of the training data is from the 21.10 version of the ParIce corpus (Barkarson & Steingrímsson, 2019; Steingrímsson & Barkarson, 2021). Finally, we used one online system, Google Translate[5]. Google Translate was chosen as it is the most popular MT system used by the general public in Iceland.

3.2 Automatic Evaluation Approaches

In order to make a general comparison of the MT systems used, we calculated BLEU and chrF++ for translations of all sentences in the dataset. The scores for BLEU[6] and chrF++[7] were calculated using Sacrebleu (Post, 2018). Sacrebleu signatures are given in footnotes and results reported in Table 1.

Google Translate scored the highest by far, and while Miðeind's system scored slightly higher than Steingrímsson's system, the difference was not statistically significant for the chrF++ scores, as calculated using the pairwise bootstrap test (Koehn, 2004).

Furthermore, we devised a simple automatic approach to gauge how well the MT systems managed to process the idiomatic expressions. Each machine-translated output is either assigned a pass or a fail. The translation gets a pass if it contains all content words of the translation in the dataset. For example, for the sentence *Sigurður fékk sér kríu*, translated in the dataset as 'Sigurður took a nap', the MT translation has to contain the words 'took' and 'nap' to receive a pass. If it does not contain both words, it is assigned a fail. The results of this approach are reported in Table 1, titled AutoIE for Automatic Idiom Evaluation. The score is given as a percentage, representing the ratio of the translated output that the approach classifies as correct.

3.3 Manual Evaluation

To assess whether our automatic approach is useful, all translations of the approx. 1,000 sentences, across all three MT systems were evaluated by a professional translator whose task was only to look at the MWE and assess whether it was translated correctly. The evaluator would select one of three options: *Correct translation*, *Incorrect translation* and *Unusual translation but can be understood*. He was only to look at the MWE and disregard all other possible errors in the translation. The results are given in Table 2.

Upon inspecting the results, we find that idioms that can be translated word by word from Icelandic into English, such as *Sigurður var úlfur í sauðargæru* ('Sigurður was a wolf in sheep's clothing') and *Sigurður bjargaði andlitinu* ('Sigurður saved face'; lit. 'Sigurður saved **the** face'), are most likely to be translated correctly. Idioms that require translating into an idiom that has the same meaning but uses a

MT System	Correct (%)	Understandable (%)	Incorrect (%)
Steingrímsson (filtered parallel data)	18.8	12.0	69.2
Miðeind (using backtranslations)	25.7	12.5	61.8
Google Translate	37.4	11.3	51.3

Table 2: Manual evaluation of the three MT systems.

[5] Google Translate was used to translate the sentences on April 22, 2024.

[6] BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

[7] chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

different metaphor are less likely to be translated correctly. Example of that could be *Sigurður er eldri en tvævetur*, literally ‘Sigurður is older than two winters old’, which would normally be translated into ‘Sigurður was not born yesterday’, or an idiom containing words where the most common sense is not the one carried in the idiom, such as *Sigurður rak lestina* (‘Sigurður trailed behind’) which contains the word *lest*, perhaps most commonly meaning a locomotive train and translated as ‘train’.

4 Future Work

In most cases, each Icelandic example is given only one translation in our dataset, although more translations may be valid. Adding additional valid translations for each example would be useful in order to use the dataset to automatically evaluate the capabilities of an MT system to translate idiomatic expressions. By comparing the translations deemed correct in the human evaluation to the translations given in the data set, we can add more valid translations. We intend to do so in order to make the data set even more viable for automatic evaluation.

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

5 Conclusions

The evaluation results, and the result analysis, indicate that available MT systems commonly fail when translating idiomatic expressions. Specialized evaluation sets, such as the one introduced in this paper, can be used to gauge the capabilities of MT systems. The simple automatic approach introduced here provides results in line with a thorough manual evaluation, indicating that it may be sufficient to help in the selection of the best system in this regard, when needed.

References

Barkarson, S., & Steingrímsson, S. (2019). Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 140–145.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. *Proceedings of the Third International Conference on Language Resources and Evaluation*, 1934–1940.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 388–395.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Popović, M. (2017). chrF++: words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, 612–618.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191.

Símonarson, H. B., Jónsson, H. P., Ragnarsson, P. O., Ingólfsdóttir, S. L., Þorsteinsson, V., & Snæbjarnarson, V. (2022). Long Context Translation Models for English-Icelandic translations (22.09) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/278

Steingrímsson, S. (2023). *Effectively compiling parallel corpora for machine translation in resource-scarce conditions* [Doctoral dissertation, Reykjavik University].

Steingrímsson, S., & Barkarson, S. (2021). ParIce: English-Icelandic parallel corpus (21.10) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/145

Úlfarsdóttir, Þ. (2014). ISLEX – a Multilingual Web Dictionary. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2820–2825.

Prepare to be Amazed: NoticIA, the Spanish Clickbait Dataset Transforming the Way We Read News

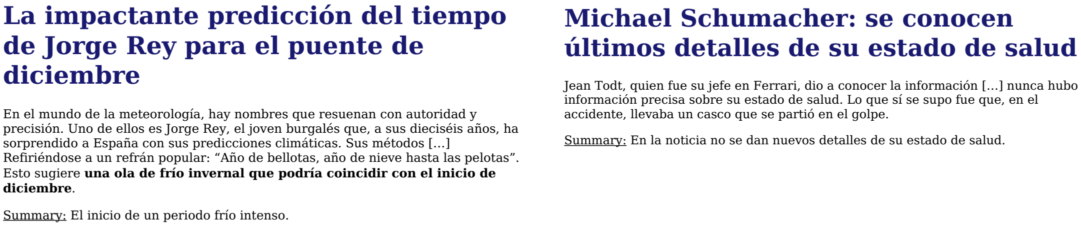
Begoña Altuna and Iker García-Ferrero
HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country UPV/EHU
Donostia, Spain
begona.altuna;iker.garciaf@ehu.eus

Abstract

NoticIA is a dataset comprising 850 Spanish news articles with clickbait headlines, each accompanied by a concise, human-generated summary sentence. This task requires sophisticated text comprehension and summarisation skills, as models must infer and connect various pieces of information to address the users' information needs prompted by the clickbait headline. We assess the Spanish text understanding abilities of several cutting-edge large language models. Furthermore, we utilise the dataset to train ClickbaitFighter, a specialised model that performs nearly as well as humans on this task.

1 Introduction

In the digital age, clickbait headlines have become a significant concern for media consumers and publishers. These sensationalised and misleading headlines lure readers to scroll through irrelevant details before finding the main idea at the expense of accurate reporting and journalistic integrity. Summarising low-quality articles with clickbait headlines is a challenging benchmark for Large Language Models (LLMs). Unlike general summarisation, this task requires models to accurately interpret the headline and extract key information hidden within filler content, testing models' ability to infer and connect diverse information to meet user needs suggested by the clickbait headline. This can be seen as an "ultrasummary" generation task, where lengthy articles are summarised into a single sentence or word. Figure 1 shows examples of clickbait headlines and human-written summaries from our dataset.



La impactante predicción del tiempo de Jorge Rey para el puente de diciembre

En el mundo de la meteorología, hay nombres que resuenan con autoridad y precisión. Uno de ellos es Jorge Rey, el joven burgalés que, a sus dieciséis años, ha sorprendido a España con sus predicciones climáticas. Sus métodos […] Refiriéndose a un refrán popular: "Año de bellotas, año de nieve hasta las pelotas". Esto sugiere **una ola de frío invernal que podría coincidir con el inicio de diciembre**.

Summary: El inicio de un periodo frío intenso.

Michael Schumacher: se conocen últimos detalles de su estado de salud

Jean Todt, quien fue su jefe en Ferrari, dio a conocer la información […] nunca hubo información precisa sobre su estado de salud. Lo que sí se supo fue que, en el accidente, llevaba un casco que se partió en el golpe.

Summary: En la noticia no se dan nuevos detalles de su estado de salud.

Figure 1: Examples of clickbait headlines. The headline is followed by a long article in which the answer to the headline is located at the end of the article.

NoticIA (NoticIA) comprises 850 Spanish news articles featuring single-sentence summaries written by humans in a headline-news body-ultra-summary triplet format, which makes a valuable high-quality text-understanding benchmark in Spanish for the evaluation of LLMs. We evaluate several state-of-the-art text-to-text Large Language Models (LLMs) in zero-shot settings using NoticIA and we also fine-tune and release LLMs trained on our dataset.

2 Related works

The Clickbait Challenge (Potthast et al., 2018) in 2017 was a milestone in the advancement of clickbait headline detection in an over-a-decade old research field. Twelve teams took part and the Webis Clickbait Corpus 2017 was released which contained 38,517 annotated tweets and that has since been used in a list of clickbait detection works, e.g. Zheng et al. (2021).

Our work, however, is more closely related to the clickbait spoiling task. The PAN Clickbait Challenge at SemEval 2023 (Fröbe et al., 2023) sought to classify the types of spoilers based on their structure and to generate spoilers according to them. 23 teams submitted their systems which were tested on the Webis Clickbait Spoiling Corpus 2022, a 5,000 headline-news-spoiler corpus collected from social media.

3 The NoticIA dataset

The NoticIA dataset comprises 850 clickbait headline, news body and summary triplets. Articles deal with a wide range of topics, all are in Spanish, and come from Spanish newspapers (83%) and Latin American newspapers (17%). Summaries were obtained manually through a task in which annotators were asked to produce the shortest possible summary that responded solely to the clickbait headline. Table 1 lists the average word counts for the headlines, article bodies, and summaries. On average, the article body was summarised with a 98% reduction.

	Train	Dev	Test	Total
Article no.	700	50	100	850
Avg. Headline Len.	16	17	17	17
Avg. Article Len.	544	663	549	552
Avg. Summary Len.	12	11	11	12

Table 1: Average word counts for the headlines, article bodies, and summaries.

The NoticIA dataset can be accessed from the CLARIN Virtual Language Observatory.[1]

4 Experiments

Our evaluation aims to shed light on the performance of the current state-of-the-art LLMs in Spanish. We evaluate a diverse set of instruction-tuned LLMs, ranging from 2 billion to 70 billion parameters: Deepseek (Bi et al., 2024), LLama2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Solar (Kim et al., 2023), Mixtral (Jiang et al., 2024), StableLM (stability.ai, 2023), Yi (01.AI et al., 2024) and Qwen (Bai et al., 2023), as well as further fine-tuned versions of these models which have improved capabilities and the GPT3.5 and GPT4 (OpenAI, 2023) commercial productions.

In addition, we fine-tune three task-specific models using the 700 training examples in the NoticIA dataset: ClickbaitFighter-10B (based on Nous-Hermes-2-SOLAR-10.7B), ClickbaitFighter-7B, (based on openchat-3.5-0106), and ClickbaitFighter-2B, (based on Gemma-2B-IT).[2]

We adopt a zero-shot prompt evaluation setup, for which we provide an instruction that describes the task and the annotation guidelines. The model receives the headline and the article body, and we expect a summary of the article as the output. As standard in summarisation tasks, we use the ROUGE score (Lin, 2004) to evaluate the summaries produced by the models.

5 Results

The results of the models are displayed in Figure 2. Task-specific models are able to produce concise summaries, which are almost twice as short as those from models in zero-shot settings. Although it is

[1] https://vlo.clarin.eu/record/oai_58_b2share.eudat.eu_58_b2rec_47_8e1473ddccde41debe8a43dba179b47c?1&q=noticia&index=0&count=67

[2] A demo of the ClickbaitFighter-7B model can be found in the SomosNLP's Hugging Face spaces: https://hf.co/spaces/somosnlp/NoticIA-demo.

a small model, ClickbaitFighter-2B performs better than any of the SotA LLMs of our experiment in the zero-shot setting, which makes it ideal for situations where computational resources are reduced. ClickbaitFighter-10B and 7B, on their side, achieve a summary quality close to the human baseline.

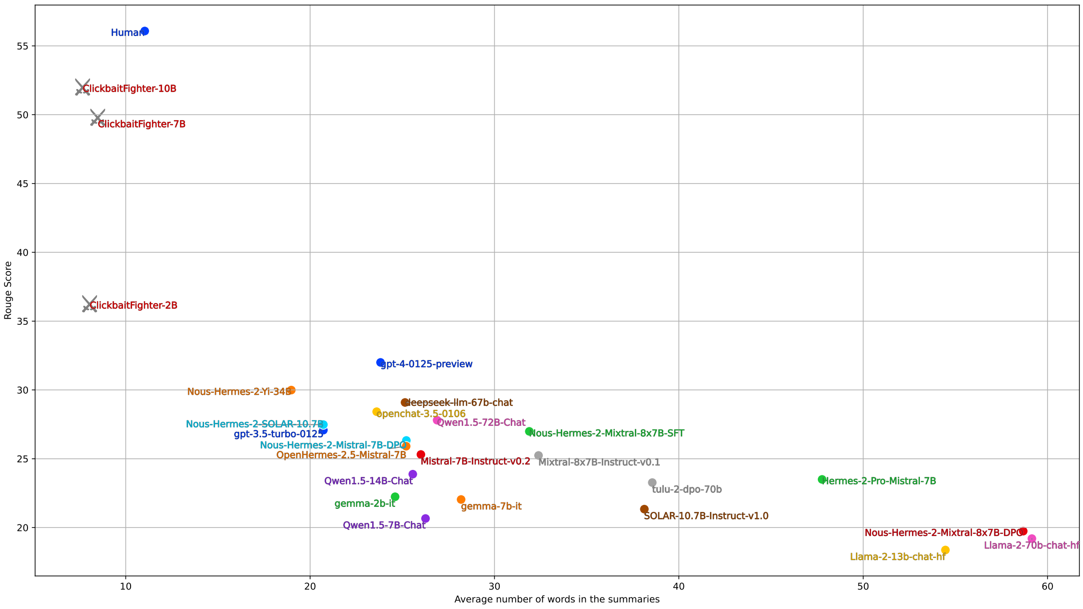


Figure 2: ROUGE score and average summary lengths for all models evaluated in our dataset. The Y-axis represents the ROUGE score, while the X-axis indicates the average number of words in the summaries. A higher ROUGE score and a shorter summary length are considered optimal.

6 Conclusions

Our experiments show that NoticIA can effectively assess the text comprehension capabilities of LLMs in Spanish. Moreover, it is also demonstrated that ultrasummarisation of clickbait news is a challenging task for the currently existing models. NoticIA also proves to be effective for training LLMs for the task of clickbait article summarisation as the performance of the fine-tuned models reflects. In the future, the dataset could further be extended, as results indicate that expanding the dataset beyond the 750 training samples could further enhance model performance. The dataset could also be easily enriched with news in other languages as clickbait journalism is a global issue, and it is not to forget that multimodality (e.g. video and text) is worth being explored.

Acknowledgments

This work has been partially funded by:

- DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

- Disargue (TED2021-130810B-C21) MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR.

- DeepKnowledge (PID2021-127777OB-C21) MCIN/AEI/10.13039/501100011033 and by FEDER, EU.

- Red CLARIAH-ES (RED2022-134527-E) of the Spanish Ministry of Science, Innovation and Universities (State Program to Promote Scientific-Technical Research and its Transfer).

References

01.AI, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., … Dai, Z. (2024). Yi: Open Foundation Models by 01.AI.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., … Zhu, T. (2023). Qwen Technical Report. *CoRR*, *abs/2309.16609*. https://doi.org/10.48550/ARXIV.2309.16609

Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., Gao, K., Gao, W., Ge, R., Guan, K., Guo, D., Guo, J., Hao, G., Hao, Z., … Zou, Y. (2024). DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *CoRR*, *abs/2401.02954*. https://doi.org/10.48550/ARXIV.2401.02954

Fröbe, M., Stein, B., Gollub, T., Hagen, M., & Potthast, M. (2023, July). SemEval-2023 Task 5: Clickbait Spoiling. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, & E. Sartori (Eds.), *Proceedings of the 17th international workshop on semantic evaluation (semeval-2023)* (pp. 2275–2286). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.312

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *CoRR*, *abs/2310.06825*. https://doi.org/10.48550/ARXIV.2310.06825

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de Las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M., Stock, P., Subramanian, S., Yang, S., … Sayed, W. E. (2024). Mixtral of Experts. *CoRR*, *abs/2401.04088*. https://doi.org/10.48550/ARXIV.2401.04088

Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J., Ahn, C., Yang, S., Lee, S., Park, H., Gim, G., Cha, M., Lee, H., & Kim, S. (2023). SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. *CoRR*, *abs/2312.15166*. https://doi.org/10.48550/ARXIV.2312.15166

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013

OpenAI. (2023). GPT-4 Technical Report. *CoRR*, *abs/2303.08774*. https://doi.org/10.48550/ARXIV.2303.08774

Potthast, M., Gollub, T., Komlossy, K., Schuster, S., Wiegmann, M., Garces Fernandez, E. P., Hagen, M., & Stein, B. (2018, August). Crowdsourcing a Large Corpus of Clickbait on Twitter. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th international conference on computational linguistics* (pp. 1498–1507). Association for Computational Linguistics. https://aclanthology.org/C18-1127

stability.ai. (2023). Introducing Stable LM Zephyr 3B: A New Addition to Stable LM, Bringing Powerful LLM Assistants to Edge Devices. https://stability.ai/news/stablelm-zephyr-3b-stability-llm

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, *abs/2307.09288*. https://doi.org/10.48550/ARXIV.2307.09288

Zheng, J., Yu, K., & Wu, X. (2021). A deep model based on Lure and Similarity for Adaptive Clickbait Detection. *Knowledge-Based Systems*, *214*, 106714. https://doi.org/https://doi.org/10.1016/j.knosys.2020.106714

On the creation of multilingual NER and ASBA workflows for literary-historical texts with chat-based LLMs

Tess Dejaeghere
Ghent University, Ghent
tess.dejaeghere@ugent.be

Pranaydeep Singh
Ghent University, Ghent
pranaydeep.singh@ugent.be

Els Lefever
Ghent University, Ghent
els.lefever@ugent.be

Julie Birkholz
Ghent University, Ghent
julie.birkholz@ugent.be

Abstract

Challenges regarding information extraction tasks from literary-historical texts arise from the unique nature of these texts, including historical languages, scarce benchmark datasets, variable annotation schemes, and digitization quality issues. Despite efforts to explore IE capabilities of open-source chat-based Large Language Models (LLMs), their application to literary-historical texts remains uncharted territory. The paper discusses the development of user-in-the-loop methodologies for applying and evaluating open-source chat-based generative models for named entity recognition and aspect-based sentiment analysis in literary-historical texts. Using a diverse corpus of travelogues spanning English, French, Dutch, and German from the 18th to the 20th century as a use-case, we present adaptable, multilingual workflows in Jupyter Notebooks. Our contributions include sharing datasets, notebooks and annotations for ABSA and NER via the CLARIN infrastructure and initiating engagement of the (digital) humanities community in assessing chat-based LLMs for information extraction from literary-historical texts.

1 Introduction

At the time of writing, discourse surrounding Large Language Models (LLMs) is reaching a fever pitch. Chat-based LLMs enable users to engage with training data using natural language, revolutionizing communication paradigms and propagating a wide adoption of AI-tools across text-based tasks. Recent efforts have been dedicated to exploring and assessing generative LLMs' performance for information extraction tasks across various linguistic spaces and domains with variable results (Han et al., 2023; Li et al., 2023; Sarmah et al., 2023; Xie et al., 2023; Xu et al., 2023).

Currently the application of LLMs as well as the evaluation of text analysis tools such as Named Entity Recognition (NER) and sentiment analysis for literary-historical text material remains limited. Literary texts pose unique challenges for annotation due to their subjective nature and stylistic properties, hindering standardization of annotation and evaluation practices across the computational literary domain (Bamman et al., 2019; Ehrmann et al., 2021; Ivanova et al., 2022; Kleymann & Stange, 2021; Rebora, 2023). The rigid nature of sentiment analysis tools, which usually procure a sentiment score on the level of the document, sentence or paragraph on a three-point scale, hardly suffices to model this multi-layered sentimental expression. Aspect-based sentiment analysis (ABSA), which evaluates sentiment on the level of specific aspects or entities may offer a more detailed and interpretable perspective on sentiment expression in literary text.

In addition, treating works of literary-historical nature as data from which to extract information comes with specific methodological and technical challenges related but not limited to the presence of historical languages, the limited availability of domain-specific benchmark datasets, the wide variability of annotation schemes, and the variable quality of both digitization processes and off-the-shelf annotation tools (McGillivray et al., 2020; Moretti, 2013). Handling historical corpora is further influenced by errors produced by Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), as well as spelling variations and concept drift (Won et al., 2018). While the lack of annotated domain-specific historical texts and hinders the development of proper NLP tool development for this text type (Kuhn & Reiter, 2015; McGillivray et al., 2020), the current high error rates of off-the-shelf annotation tools when applied to historical textual data, as well as the required level of programming expected from end users are often times intimidating to (digital) humanists – nudging them back to a familiar praxis of close reading and manual analysis (Blevins & Robichaud, 2011; Kuhn, 2019).

Fostering an understanding of these chat-based LLMs, their inner workings and potential risks, is crucial for the humanities community, given their increasing role in education and research. To support the implementation of such tools in practice, we developed user-in-the-loop methodologies in Jupyter Notebooks with open-source chat-based generative models for NER and ABSA for literary-historical texts which can be tailored to specific domains in zero- and few-shot settings, and evaluated without extensive training data. We add code for performing a quantitative evaluation by calculating evaluation metrics (F1) as well as a qualitative evaluation, by sampling some results which can be used by the scholar to adapt the prompt.

2 Data

Our workflows are built using a corpus of travelogues featuring diverse genres such as nature writing, travel memoirs, journals, and poetry. Our travel texts were sourced from various online repositories and resulted in a dataset of 3,320 texts across the languages English, French, Dutch and German as shown in Table 1.

Language	**18thC**	**19thC**	**20thC**	**Total**
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**

Table 1: Overview of languages contained in the travelogues corpus

The Jupyter Notebooks are developed using 1) a dataset of 58 texts (approx. 5000 tokens/text) across these languages, manually labelled with aspects and their sentiment scores and 2) a manually labelled dataset of 128 texts (approx. 500 tokens/text) with named entities to showcase the evaluation approach.

3 Methodology

As shown in 1, five Jupyter Notebooks are created for entity and/or aspect extraction and ABSA respectively. Three notebooks showcase the extraction of entities/aspects using 1) the open-source chat-based LLM mistralai/Mixtral-8x7B-Instruct-v0.1 in a zero-shot and few-shot prompting setting, 2) the NLP-package Flair (Akbik et al., 2019) and 3) the NLP-package spaCy (Montani et al., 2023). These notebooks take raw text as input, chunk and label the texts in an output which results in a .csv-format. This output can then be fed into the ABSA-notebooks. The ABSA-notebooks include 1) a generative workflow for zero-shot and few-shot sentiment analysis based on the Mixtral-8x7b-Instruct model and 2) if sufficient annotated data is available, the user can also choose to use our machine learning-based training pipelines developed for the languages under consideration, which extract embeddings from language-specific models available on HuggingFace to serve as input for diverse machine learning classification architectures, including SVM, AdaBoost, Random Forest, and MLP classifiers.

Evaluations are carried out quantitatively by, if gold standard annotations are available, calculating NLP-native span evaluation metrics such as F1. To encourage the exploration of qualitative evaluation methods for generative models, we introduced a sampling methodology which presents the user with random labelled entries from the test set, which will allow for the adaptation of the prompt or input data.

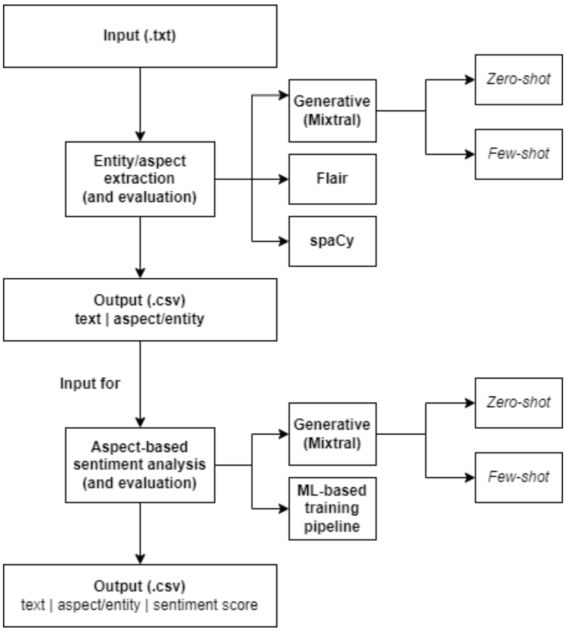

Input (.txt)
Entity/aspect extraction (and evaluation)
Generative (Mixtral)
Zero-shot
Few-shot
Flair
spaCy
Output (.csv) text | aspect/entity
Input for
Aspect-based sentiment analysis (and evaluation)
Generative (Mixtral)
Zero-shot
Few-shot
ML-based training pipeline
Output (.csv) text | aspect/entity | sentiment score

Figure 1: Overview of the notebook workflow for the NER and ABSA tasks respectively

4 Expected contributions to the CLARIN infrastructure

The notebooks, datasets and annotations for both the ABSA and NER tasks will be made available through the CLARIN infrastructure and through the open science ecosystem via CLARIAH-VL (https://clariahvl.hypotheses.org/) and also shared as tools as part of the H2020 CLS INFRA – Computational Literary Studies Infrastructure Project (https://clsinfra.io/). The output can be freely adapted and used for training and research purposes in the Digital Humanities community and beyond. Our work represents a valuable step in the direction of user-in-the-loop workflows for lesser-resourced historical languages, and a pioneering effort in the exploration of open-source chat-based LLMs for NER and ABSA applied to the literary-historical domain.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In W. Ammar, A. Louis, & N. Mostafazadeh (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics (demonstrations)* (pp. 54–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010

Bamman, D., Popat, S., & Shen, S. (2019, June). An annotated dataset of literary entities. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2138–2144). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1220

Blevins, C., & Robichaud, A. (2011). 2: A Brief History ≫ Tooling Up for Digital Humanities. Retrieved December 7, 2021, from http://toolingup.stanford.edu/?page_id=197

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021). Named Entity Recognition and Classification on Historical Documents: A Survey [arXiv: 2109.11406]. *arXiv:2109.11406*

[cs]. Retrieved February 22, 2022, from http://arxiv.org/abs/2109.11406
Comment: 39 pages.

Han, R., Peng, T., Yang, C., Wang, B., Liu, L., & Wan, X. (2023, May). Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors [arXiv:2305.14450 [cs]]. Retrieved March 12, 2024, from http://arxiv.org/abs/2305.14450
Comment: 23 pages, version 1.0.

Ivanova, R., van Erp, M., & Kirrane, S. (2022, June). Comparing Annotated Datasets for Named Entity Recognition in English Literature. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3788–3797). European Language Resources Association. Retrieved February 5, 2024, from https://aclanthology.org/2022.lrec-1.404

Kleymann, R., & Stange, J.-E. (2021). Towards Hermeneutic Visualization in Digital Literary Studies. *DHQ: Digital Humanities Quarterly*, *2*(15). Retrieved December 7, 2021, from http://digitalhumanities.org:8081/dhq/vol/15/2/000547/000547.html

Kuhn, J. (2019). Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation*, *53*(4), 565–602. https://doi.org/10.1007/s10579-019-09459-3

Kuhn, J., & Reiter, N. (2015, June). *A Plea for a Method-Driven Agenda in the Digital Humanities*.

Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. (2023, April). Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness [arXiv:2304.11633 [cs]]. https://doi.org/10.48550/arXiv.2304.11633

McGillivray, B., Poibeau, T., & Fabo, P. R. (2020). Digital Humanities and Natural Language Processing: Je t'aime... Moi non plus. *Digital Humanities Quarterly*, *014*(2). https://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Landeghem, S. V., & Peters, H. (2023, October). Explosion/spaCy: V3.7.2: Fixes for APIs and requirements. https://doi.org/10.5281/ZENODO.1212303

Moretti, F. (2013). *Distant reading*. Verso.
Modern European literature: a geographical sketch – Conjectures on world literature – The slaughterhouse of literature – Planet Hollywood – More conjectures – Evolution, world-systems, Weltliteratur – The end of the beginning: a reply to Christopher Prendergast – The novel: history and theory – Style, Inc.: reflections on 7,000 titles (British novels, 1740-1850) – Network theory, plot analysis.

Rebora, S. (2023). Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, *017*(2).

Sarmah, B., Zhu, T., Mehta, D., & Pasquali, S. (2023, October). Towards reducing hallucination in extracting information from financial reports using Large Language Models [arXiv:2310.10760 [cs, q-fin, stat]]. Retrieved March 13, 2024, from http://arxiv.org/abs/2310.10760
Comment: 4 pages + references. Accepted for publication in Workshop on Generative AI at the 3rd International Conference on AI-ML Systems 2023, Bengaluru, India.

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, *5*, 2. https://doi.org/10.3389/fdigh.2018.00002

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., & Wang, H. (2023, October). Empirical Study of Zero-Shot NER with ChatGPT [arXiv:2310.10035 [cs]]. https://doi.org/10.48550/arXiv.2310.10035
Comment: Accepted to EMNLP 2023 (Main Conference).

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., & Chen, E. (2023, December). Large Language Models for Generative Information Extraction: A Survey [arXiv:2312.17617 [cs]]. https://doi.org/10.48550/arXiv.2312.17617

Word Rain as a Service

Magnus Ahltorp
Language Council of Sweden
Institute for Language and Folklore
Stockholm, Sweden
magnus.ahltorp@isof.se

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

Abstract

Word Rain is a novel approach to the classic word cloud that uses word embeddings to make it useful for exploring the word content of a text or corpus. Downloading and running the code can, however, be prohibitively difficult or cumbersome for non-technical users and casual evaluation. Since Word Rain also requires a word embeddings model, the inexperienced or casual user would benefit greatly from a streamlined interface. We have therefore collected everything that is needed in a web based service and are making it available as a SWELANG K-centre resource.

1 Introduction

Word clouds are widely used when informally illustrating the terms in a text or corpus, most likely because of their usefulness as a graphic design element and the prevalence of easy-to-use tools. Traditional word clouds are, however, also criticised because of the lack of semantic relevance in the placement of words (Barth et al., 2014). The reader can therefore be misled into thinking that there is a relationship between nearby words where there is none. The size of the text can also be misleading, since longer words take up more space, and therefore seem more important than shorter ones.

This project aims to direct particularly researchers, but also people in general, to the novel and more useful Word Rain technique by streamlining the process to the point where it is as easy to generate as a word cloud is today.

2 Word Rain

Word Rain (Skeppstedt et al., 2024) proposes a solution to the problems with word clouds by using word embeddings reduced to one dimension as the x-axis, and word prominence as the primary basis for positioning the words on the y-axis. Several Word Rains can be generated with the same x-axis, making it possible to compare corpora.

As an example of this, in figure 1, we have generated three different word rains with the top 500 words/bigrams: the upper from the EuroParl-UdS corpus (Karakanta et al., 2018), spanning the years 1999-2017, the middle (House of Commons) and lower (House of Lords) from the British part of the ParlaMint corpus (Erjavec et al., 2023), only for the year 2017.

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

On the other hand, we can look at a very prominent word in the House of Lords corpus: *noble*. If we look in the European Parliament and House of Commons word rains, we will not find this word among the top 500 visualised. We will, however, find words that are used similarly nearby, such as *hon* (short for Honourable, title of member of parliament) in the House of Commons, and *mr* in the European Parliament.

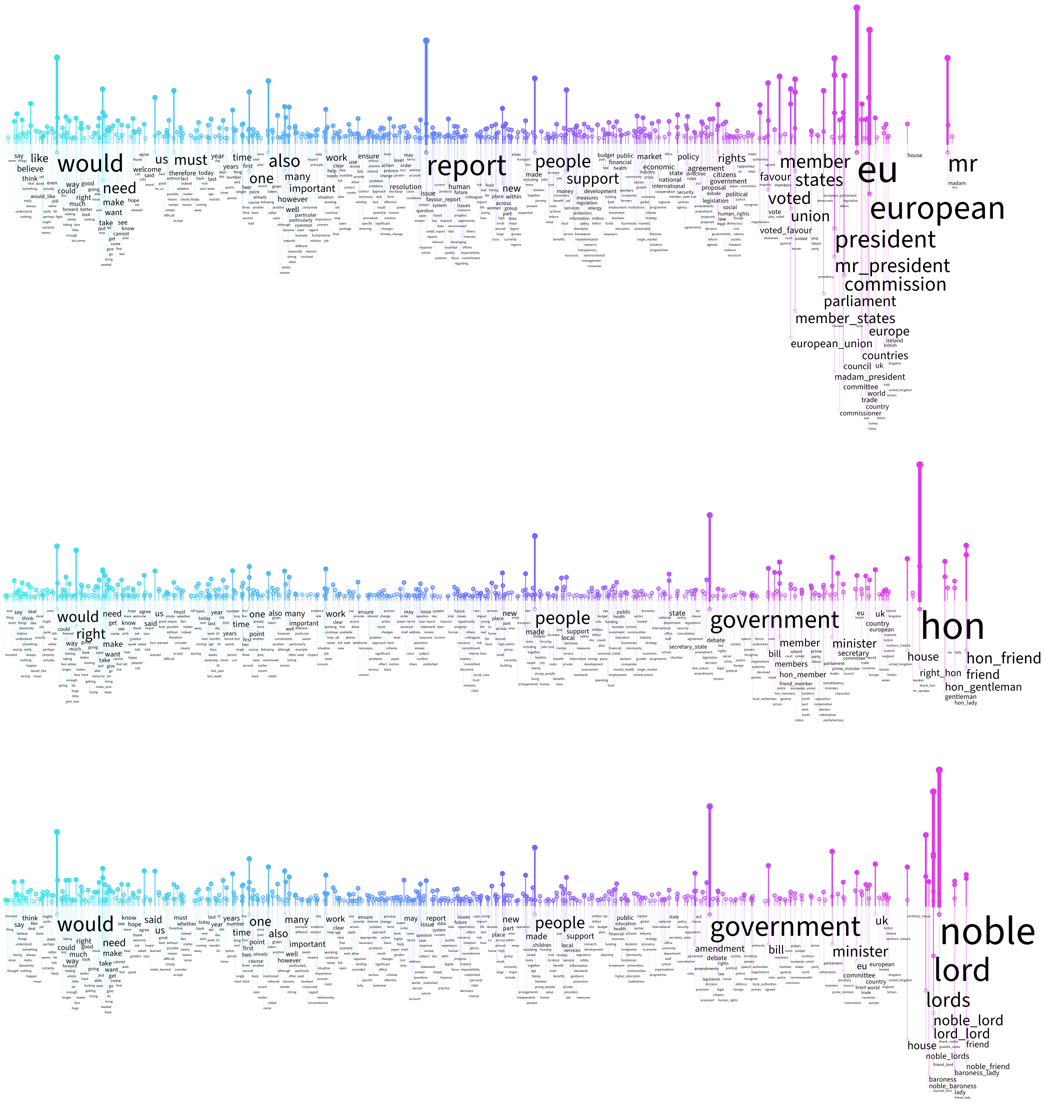


Figure 1: Example Word rains comparing corpora by using the same x-axis. The texts are from the European Parliament (EU), the House of Commons (UK) and the House of Lords (UK).

The coloured bars above the words give a sense of where there are prominent words, without the bias of word length. This makes it easy to see the peak in the European Parliament data near the far right side of the graph containing words associated with the EU such as *EU*, *commission*, and *european*, with the corresponding area having low bar heights in the UK data.

Together, these properties of Word Rain means that it can be meaningful to dig deeper, and use the graph for actual research, as opposed to traditional word clouds, where the information content is not greater than a simple sorted list of words. Skeppstedt et al. (2024) showcase the method on text comparison and lexicon development tasks and have also performed a user study.

An important observation is that the goal of the Word Rain visualisation technique is not to represent the original word embeddings in as much detail as possible, but to use the word embeddings to order the words in a meaningful way.

3 The current situation

Word clouds can currently be generated by a large number of tools, from numerous online websites to software libraries that can be easily integrated into custom solutions. Word Rain, on the other hand, requires the user to download the code, install relevant packages, find a suitable language model, and finally write a small program using the library. For the more advanced applications of Word Rain, such as comparisons between different corpora, the user would also have to manually merge the resulting graphs.

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

4 Web service

Our solution to this is to collect everything that is needed in a web based service where the user can just upload one or more text documents and choose parameters for the visualisation. The service is both available as a web site at https://wordrain.isof.se/ and as open source code that can be easily deployed in standard web server environments[1].

At the moment we have language models for Swedish, English, Finnish and Yiddish. As a next step, we are planning on adding support for the Swedish national minority language Meänkieli.

Apart from the selection of a language model, the parameters can be grouped into two categories: parameters controlling the text processing, and parameters controlling the graphical presentation. Examples of text processing parameters are whether to use inverse document frequency and/or a background corpus for prominence calculations, the maximum size of n-grams that should be treated as one term, and the desired number of words to display.

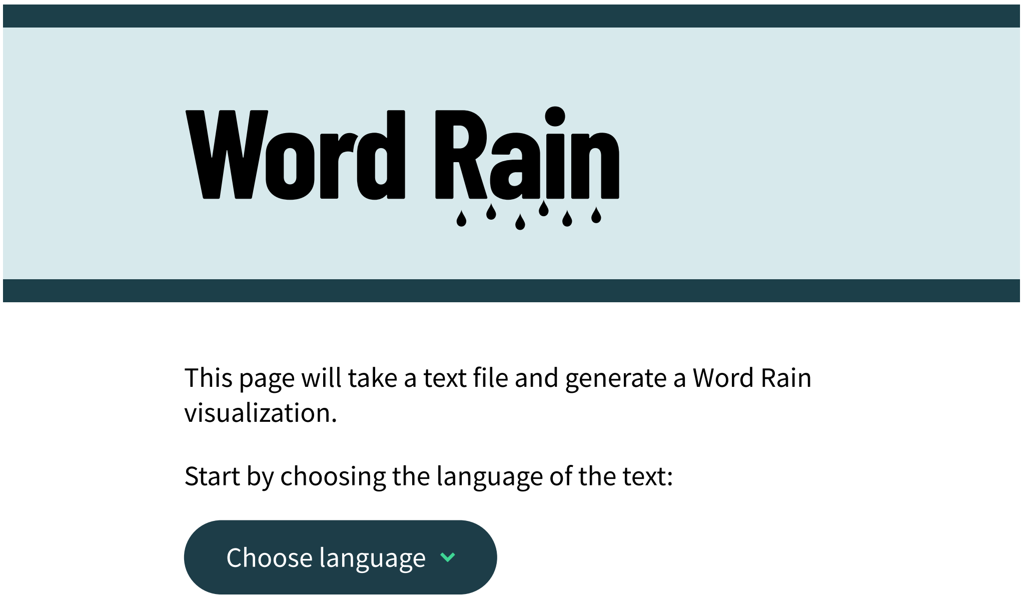
Graphical presentation parameters include how much space to dedicate to the vertical bars, and how sharp the drop-off in font size should be.

The parameters mentioned above are all configurable in the initial version, with more to come at a later stage.

5 Conclusions

Providing an easy-to-use tool for generating word rains is crucial to widespread adoption. Many potential users do not have the technical skills that are required to run the Word Rain system in its present form. Others do not have the time to evaluate a new tool even if they would eventually want to set it up themselves, should they decide to use it in their workflow. The tool is a welcome addition to the SWE-LANG K-centre repertoire, especially for the K-centre's focus: the languages of Sweden. Offering it for English as well from the start makes it useful for prototyping Word Rain use from the whole CLARIN user community.

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



Word Rain

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

Start by choosing the language of the text:

Figure 2: The Word Rain service web site at https://wordrain.isof.se/.

References

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

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., . . . Fišer, D. (2023). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1859

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

Skeppstedt, M., Ahltorp, M., Kucher, K., & Lindström, M. (2024). From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*. https://doi.org/10.1177/14738716241236188

Language Technology Initiative — Bridging the Gap between Research and Education

Inguna Skadiņa
Institute of Mathematics and Computer Science
University of Latvia, Latvia
inguna.skadina@lumii.lv

Jana Kuzmina
University of Latvia
Riga, Latvia
jana.kuzmina@lu.lv

Sergejs Kruks
Riga Stradiņš University, Latvia
sergei.kruk@gmail.com

Marina Platonova
Riga Technical University, Latvia
marina.platonova@rtu.lv

Tatjana Smirnova
Riga Technical University
Latvia
tatjana.smirnova@rtu.lv

Ilze Auziņa
Institute of Mathematics and Computer Science
University of Latvia, Latvia
ilze.auzina@lumii.lv

Abstract

In this abstract, we introduce a Language Technology Initiative — a project aiming to bridge the gap between research and education in the field of Language Technology. The key activities of the Language Technology Initiative are (1) the development of Language Technology and Computation Linguistic modules and courses for three main universities in Latvia, (2) the advancement of research and development activities in five important LT directions – language resources for digital humanities research and technology development, large language models, speech technologies, human-computer interaction and machine translation and (3) support for sustainable and open research and education through CLARIN research infrastructure. In this abstract, we highlight the most important achievements of the first project year —- developed study modules and courses. We also outline the first outcomes from the research activities and demonstrate the first synergies between research and education and the role of CLARIN infrastructure.

1 Introduction

Language resources and tools that support the Latvian language in the digital world have been developing for several decades (Skadina et al., 2022). Many of them are registered in CLARIN-LV[1] repository, are available through different on-line services, or offered by Latvian and global companies. While different Latvian and multilingual language resources and tools are widely used by the Latvian language speakers and learners, only some courses related to Language Technologies (LT), computational linguistics, and digital humanities (DH) have been included in Latvian university programs.

The need for designated Language Technology and computational linguistics modules has been recognized by the Ministry of Education and Sciences of Latvia proposing to invest in the development of high-level digital skills in three areas: Language Technology, quantum computing, and HPC through the Latvian Recovery and Resilience Plan.[2] The objective of this investment is to significantly increase the number of specialists with high-level digital skills (DESI levels 3–5) by 2026. The Language Technology Initiative (LTI) project was initiated to implement this objective through the synergy between research and education sectors.

In this abstract, we introduce the LTI project and highlight the most important achievements from the first project year. We demonstrate synergy between new study modules and the first outcomes from the

[1] https://repository.clarin.lv/repository/

[2] https://commission.europa.eu/business-economy-euro/economic-recovery/recovery-and-resilience-facility/latvias-recovery-and-resilience-plan/latvias-recovery-and-resilience-supported-projects-nation-wide-investment-scheme_en

research activities. We also describe the role of CLARIN in this project as a platform to support studies and experimentation.

2 Language Technology Initiative Project

The Language Technology Initiative project[3] aims to create a synergy between higher education, science, and industry. The project is envisioned to prepare a curriculum for Language Technology teaching, advance language resources, and create platforms and tools for studying and experimentation, as well as to conduct research involving young researchers. To reach the primary goal of the project — significantly increase the number of specialists with high-level digital skills by 2026 — the project aims to develop at least five study modules in Language Technology and educate at least 820 students (both, LT developers and users). Besides students, 12 young specialists, who can teach LT at universities, will be educated. Finally, research results from the project will be integrated into study modules and summarized in at least 10 high-level publications.

The project consortium comprises three leading Universities of Latvia — University of Latvia (UL), Riga Technical University, and Riga Stradiņš University, two research organizations — the Institute of Mathematics and Computer Science, UL (coordinator of CLARIN-LV) and the Institute of Literature, Folklore and Arts, UL, and the language technology company Tilde.

3 Education Activities

Three partner universities (University of Latvia, Riga Technical University, and Riga Stradiņš University) are involved in the preparation of the LT and computational linguistics modules and courses.

3.1 University of Latvia

At the University of Latvia both the Faculty of Science and Technology and the Faculty of Humanities are involved in the development of study modules.

3.1.1 Faculty of Science and Technology

The Faculty of Science and Technology (former Faculty of Computing) is developing two study modules — one for Bachelor students in Computer Science and another for Master Students in Computer Science. The Bachelor module was developed in 2023 and its implementation started in the Spring semester of 2024. It comprises three courses (3 ECTS for each course) — Fundamentals of Natural Language Processing, Fundamentals of Deep Machine Learning, and Introduction to Python programming language. Study materials for each course is available from Moodle platform of UL.[4]

"Fundamentals of Natural Language Processing" course aims to introduce students to the Language Technology and its use in practical applications. The course covers the basic methods as well as the most important innovations and trends in the LT field. This includes language processing and modeling at different levels of text analysis by applying both knowledge-based and data-driven approaches. The main focus is on data-driven methods and the language resources they require. CLARIN language resources and tools, in particular from the CLARIN-LV repository, are widely used in this course. For example, Latvian processing pipeline NLP-pipe[5] is used for text analysis, while different corpora, e.g., Rainis corpus[6] and annotated Latvian data from the "Full Stack of Latvian Language Resources for NLU" dataset[7], are used for rule-based and data-driven approaches.

The "Fundamentals of Deep Machine Learning" course aims to provide an overview of modern applications of machine learning and develop practical skills in using deep neural networks for common machine learning tasks — classification, text and image processing. The objective of this course is to provide an introduction into artificial neural network based models and to introduce to existing API frameworks for training such models.

[3] https://www.vti.lu.lv/en/

[4] https://estudijas.lu.lv/

[5] http://hdl.handle.net/20.500.12574/4

[6] http://hdl.handle.net/20.500.12574/41

[7] https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/5

3.1.2 Faculty of Humanities

The Faculty of Humanities has developed four modules covering all educational levels, providing a module for Doctoral students in Linguistics, a module for Master students of English Studies, a module for Master students of Latvian Language, Literature and Culture as well as several courses for Bachelor students of Latvian Language and English, European Languages and Business Studies. The learning outcomes of the modules comprise various digital competence areas, i.e. information and data literacy, communication and collaboration, digital content creation, safety and problem-solving. The implementation and approbation of four courses from the modules started in February 2024.

The module for Doctoral students in Linguistics further develops the academic competence for conducting independent and innovative research in the LT field. It consists of two courses: Corpus Linguistics in the Context of Digital Humanities and Language, Thinking and Language Acquisition. The module for Master students of English Studies provides in-depth knowledge, skills and competence for LT use in English linguistics and literature studies. It consists of three courses: Corpus Linguistics, Programming Languages for Linguists and Traditional and Electronic Lexicography. The module for Master students of Latvian Language, Literature and Culture provides in-depth knowledge, skills and competence for LT use in Latvian linguistics and literature studies. It comprises two courses: Morphemics, Morphonology and Latvian Language Morpheme Database and Spoken Data Processing and Analysis.

The modules bring to the foreground the comparison of artificial and human intelligence of language structure, the research on thinking strategy role in language acquisition, evaluating the efficiency of digital language learning platforms and applications, versatile data analysis, design and use of language corpora in the context of digital humanities and various linguistic sub-branches. They also consider programming solutions for professional and research activities in the context of text analysis, analysis of electronic dictionaries and their content critical evaluation, the study of Latvian language morphemes and creating the database of word-building models, technologies of extracting audio and video materials, spoken text recognition, transcription and extracted data analysis, automated text analysis in research and professional discourse, qualitative data automated analysis, e.g. thematic or sentiment analysis, data visualisation and creating interactive narratives. It considers the use of different digital resources and tools, including ones provided by CLARIN consortium — *Sketch Engine, english-corpora.org, Iteman*, ELAN, NLP-PIPE, Praat, WebCorp, CLAWS, Wmatrix, LancsBox as well as *NLTK* and *SpacY libraries*, *NVivo*, *Praat*, *Voyant* Tools, World Atlas of Language Structures, Knightlab, Miro.

3.2 Riga Technical University

The Faculty of Computer Science, Information Technology and Energy of Riga Technical University is committed to developing two modules: Language Technologies for Multimodal Information Processing and Language Technologies for Translation Studies.

The module "Language Technologies for Multimodal Information Processing" is primarily aimed at developing comprehensive and highly specialized knowledge, competencies, and skills of the students undertaking study programs in humanities, interdisciplinary STEM+, and information technology whereas the module "Language Technologies for Translation Studies" covers a wide range of NLP applications in such areas as language service provision, edutainment, and processing of parallel corpora. Both modules comprise 6 study courses in total: Digital Semantics, Multimodal Digital Semiotics, Digital Sentiment Analysis, Digital Edutainment Elements in Translation, Machine Learning for Textual Data Processing, and Machine Translation and Localization. Students will develop such skills as data mining, processing and analysis, speech tagging, emotion detection, and sentiment analysis, skills in working with the Python programming language and its libraries, sentiment analysis models, and various visualization tools.

In order to reach the widest possible audience, popularizing and expanding the environment for the use of the national language (through the prism of a foreign language) and encouraging its use in the contrastive perspective, modules are being launched and implemented in a remote format as MOOCs at the dedicated platform of Riga Technical University.[8]

[8] https://mooc.rtu.lv/

3.3 Riga Stradiņš University

The Department of Social Sciences of the Riga Stradiņš University has developed two postgraduate courses. The Discourse Analysis course introduces basic methods of computer-assisted analysis of texts and prompts students to use a corpus-driven approach to formulate their research questions. Students learn to find and interpret concordances and collocations in Latvian language corpora from the Latvian National Corpora Collection (Saulite et al., 2022). [9] When learning basic Bash commands, students will be able to create mini corpora of their interest.

Discourse Analysis inclines to a qualitative methodology that involves little computation while the second course, Corpus Analysis, develops quantitative research methodology skills. Students learn to use *Sketch Engine* to extract data from the on-line Latvian corpora and to write Bash commands to work with their own mini corpora. In addition, students are encouraged to conduct comparative studies working with multilingual corpora, e.g., ParlaMint[10], available at the CLARIN-SI repository. The course explains the main mathematical operations and the essence of statistical measures. To raise interest in corpus statistics and to explain the meaning behind the numbers, the course relies on corpora of Latvian literature. Statistics of pronouns, names of colours, and grammatical forms vividly demonstrate the diachronic, gender, and generic (prose vs. poetry) regularities of language use. Particular attention is given to the overuse of ambiguous grammatical forms in genres striving for objectivity: news media, law, policy documents, and Ph.D. dissertations. Both courses are taught in hybrid form: video lectures, tests, several mini corpora, and calculators are available on the Moodle learning platform; tutorials are held in real-time mode on-line or in class. Students are awarded 3 ECTS for each course.

4 Research and Development

The research and development activities in the Language Technology Initiative are organized around five key topics: development of language resources for speech and text processing, creation of large pre-trained language models, advancement of state-of-the-art speech technologies, machine translation and development of software platforms and a shared technical infrastructure for education and research.

Several key resources, such as *Tezaurs.lv* dictionary, Latvian National Corpora Collection (LNCC), and Latvian Treebank, are being extended and regularly updated in CLARIN-LV repository and used in different courses described above. Also, Machine Translation tools developed through project are used in educational activities and by individual students during their work on thesis or course papers.

A completely new resource under development is Latvian Common Voice corpus. Whereas a year ago, the majority of the Latvian speech corpora, used by research institutions and language technology companies, were not open and freely available, it was decided to collect diverse Latvian speech samples and their transcriptions via crowdsourcing activity Balsu talka.[11] During this crowdsourcing initiative the size and speaker diversity of the Latvian Common Voice corpus has increased tenfold in less than a year (Dargis et al., 2024). The goal of this ongoing initiative is not only to enlarge the corpus, but also to make it more diverse in terms of speakers and accents, intonations, grammar and lexicon. The latest release of the BalsuTalka Speech corpus was made available for linguistic analysis as part of the LNCC.[12] A successful follow-up initiative was also launched for Latgalian, which has been recognized as an endangered historic variant of Latvian with 150k speakers.

Conclusions, Results and Next Steps

In this abstract, we introduced the Language Technology Initiative project and highlighted the main achievements from the project's first year, during which the initial set of modules and courses were developed. The implementation of these modules and courses started in the Spring semester of 2024[13] and 339 students successfully completed at least one course: 241 at the University of Latvia (93 students

[9] https://korpuss.lv/

[10] http://hdl.handle.net/11356/1859

[11] https://balsutalka.lv

[12] https://korpuss.lv/en/id/BalsuTalka

[13] Information about the actual courses are published at the LTI website https://www.vti.lu.lv/en/education/

were from computer science, while 148 students were from humanities) and 98 students from the Riga Stradiņš University. Different language resources and tools from the CLARIN-LV and other CLARIN repositories were used in developed courses. Since language technologies are a dynamic field, we plan to continue developing and improving course materials during the next years. In particular, concentrating on the inclusion of outcomes from the research activities into relevant course materials. By the end of the project, we also plan to register developed modules and courses in DH Course Registry and to contribute developed course materials to the CLARIN Learning Hub.

Acknowledgments

This work has been supported by the ”Language Technology Initiative” project (No. 2.3.1.1.i.0/1/22/I/CFLA/002), funded by the European Union Recovery and Resilience Mechanism Investment and the National Development Plan.

References

Dargis, R., Znotins, A., Auzina, I., Saulite, B., Reinsone, S., Dejus, R., Klavinska, A., & Gruzitis, N. (2024). Balsutalka.lv – boosting the common voice corpus for low-resource languages. *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2080–2085.

Saulite, B., Darģis, R., Gruzitis, N., Auzina, I., Levane-Petrova, K., Pretkalnina, L., Rituma, L., & et al. (2022). Latvian national corpora collection – korpuss.lv. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5123–5129.

Skadina, I., Saulite, B., Auzina, I., Gruzitis, N., Vasiljevs, A., Skadins, R., & Pinnis, M. (2022). Latvian language in the digital age: The main achievements in the last decade. *Baltic Journal of Modern Computing*, *10*(3), 490–503.

CLASSLA-Express: a Train of CLARIN.SI Workshops on Language Resources and Tools with Easily Expanding Route

Nikola Ljubešić and Taja Kuzman
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
{nikola.ljubesic,taja.kuzman}@ijs.si

Ivana Filipović Petrović
Linguistic Research Institute of the
Croatian Academy of Sciences and Arts
Zagreb, Croatia
ifilipovic@hazu.hr

Jelena Parizoska
Faculty of Teacher Education
University of Zagreb, Croatia
jelena.parizoska@ufzg.unizg.hr

Petya Osenova
AI and Language Technologies Division
IICT-BAS
Sofia, Bulgaria
petya@bultreebank.org

Abstract

This paper introduces the CLASSLA-Express workshop series as an innovative approach to disseminating linguistic resources and infrastructure provided by the CLASSLA Knowledge Centre for South Slavic languages and the Slovenian CLARIN.SI infrastructure. The workshop series employs two key strategies: (1) conducting workshops directly in countries with interested audiences, and (2) designing the series for easy expansion to new venues. The first iteration of the CLASSLA-Express workshop series encompasses 6 workshops in 5 countries. Its goal is to share knowledge on the use of corpus querying tools, as well as the recently-released CLASSLA-web corpora – the largest general corpora for South Slavic languages. In the paper, we present the design of the workshop series, its current scope and the effortless extensions of the workshop to new venues that are already in sight.

1 Introduction

We present the CLASSLA-Express workshop series as an approach to promote CLARIN.SI[1] linguistic resources and its infrastructure. This is achieved by 1) conducting workshops in countries with interested audiences, eliminating the need for individuals to travel to specific venues; 2) designing the workshops to be easily scalable to additional locations. By directly bringing workshops to countries with interested audiences and in some venues conducting the workshops in the national languages, a broader and more diverse group of individuals beyond the existing CLARIN.SI research community can be reached, including those who may face constraints in traveling or have a lower proficiency in English. This approach is not only more inclusive but also environmentally sustainable by only requiring the workshop lecturers to travel. Furthermore, the workshop series is structured in a way that facilitates seamless expansion to new locations. We share teaching materials, available in multiple languages, with interested colleagues from other institutions and nations, enabling the effortless extension of the workshop series to communities that were not initially targeted in the first iteration of CLASSLA-Express. Following the announcement of the first six workshops, researchers from Montenegro, Bosnia and Herzegovina, Croatia and Serbia have already expressed interest to take on the workshop, and the extension of the workshop series to additional countries and cities is already in sight.

The first iteration of the CLASSLA-Express workshop series comprises six workshops that took place from April to September 2024 in the following cities: Zagreb (Croatia), Rijeka (Croatia), Belgrade (Serbia), Skopje (North Macedonia), Sofia (Bulgaria)[2], and Ljubljana (Slovenia). The main goal of the work-

[1] https://www.clarin.si/info/about/

[2] The workshop in Sofia is also supported by CLaDA-BG: https://clada-bg.eu/en/

shop is to familiarize the participants with the CLASSLA-web corpora (Ljubešić & Kuzman, 2024), the largest general corpora for South Slavic languages, as well as with other corpora and tools provided by the CLARIN.SI infrastructure, including the pipeline for linguistic annotation CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023) and concordancers (corpus querying tools). Of the six workshops, three are aimed at students of South Slavic languages, linguists and university lecturers and are held at universities which have Slavic departments. The other three workshops are held at international conferences on lexicography, language technologies and digital humanities. Those are aimed at linguists, lexicographers, computational linguists, corpus linguists and digital humanities scholars.

In this extended abstract, we present the workshop scope and the topics it covers (Section 2), the first iteration of the CLASSLA-Express workshop series (Section 3), the promotional strategies employed to engage workshop participants, such as enhancing brand visibility and disseminating information through various channels (Section 4) and the workshop framework designed to facilitate expansion to additional locations (Section 5).

2 Goal and Content of the Workshops

The CLASSLA-Express series of workshops aims to show participants how to use the CLASSLA-web corpora (Ljubešić & Kuzman, 2024) in language research, but also language teaching, designing corpus-informed dictionaries and grammars, and in digital humanities. The CLASSLA-web corpora cover Slovenian, Croatian, Bosnian, Montenegrin, Serbian, Macedonian, and Bulgarian, which makes them the first collection of comparable corpora that cover the entire language group. The corpus collection comprises a total of 13 billion tokens of texts from 26 million documents and represents ones of the biggest openly-available corpora for each language. What is more, the Macedonian CLASSLA-web corpus is the first linguistically annotated corpus available for this language. The CLASSLA-web corpora are available for download at the CLARIN.SI repository[3], and for querying at the CLARIN.SI NoSketch Engine concordancer[4]. They are characterized by their recency, substantial size, and diverse composition encompassing various genres, registers, and topics. Consequently, they represent a highly valuable resource for conducting language analyses, developing lexicons, and facilitating other forms of linguistic research.

The main part of the workshop comprises hands-on exercises showing how to create queries in Croatian (Ljubešić et al., 2024b), Macedonian (Ljubešić et al., 2024c), Serbian (Ljubešić et al., 2024d), Bulgarian (Ljubešić et al., 2024a), and Slovenian (Ljubešić et al., 2024e) web corpora to obtain data on meanings and uses of words, word forms, collocations and grammatical patterns. The presented queries were carefully designed to cater for participants with various research interests: morphology (word forms), grammar (e.g., case, aspect, sentence patterns), lexicology (collocations, idiomatic expressions) and discourse analysis (uses of words and constructions in different types of texts).

In addition, as many corpus linguistic analyses are based on linguistic annotations of the corpora, e.g., part-of-speech tags, we familiarize the participants with the automatic language processing tool CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023) which was used to annotate the CLASSLA-web corpora. The tool is also accessible through a web platform[5] which allows researchers to explore it and evaluate its limitations that should be taken into account when conducting a linguistic study of automatically annotated data. Lastly, the CLASSLA-Express workshops introduce the participants to a wide range of language resources and technologies that are provided for their languages by the CLARIN.SI infrastructure, as well as the knowledge-sharing initiatives and support services offered by the CLARIN Knowledge Centre for South Slavic languages (CLASSLA)[6].

3 First Iteration of the CLASSLA-Express Workshop Series

The first iteration of the CLASSLA-Express workshops took place from April to September 2024 across five countries and six cities. These workshops were designed as half-day sessions, lasting approximately

[3] https://www.clarin.si/repository/xmlui/

[4] https://www.clarin.si/ske/

[5] https://clarin.si/oznacevalnik/eng/

[6] https://www.clarin.si/info/k-centre/

four hours each. Table 1 provides information on the dates and locations of the workshops. The workshops were conducted in Croatian in Zagreb, Rijeka and Skopje, Serbian in Belgrade, Bulgarian in Sofia, and English in Ljubljana.

Each workshop was intended for approximately 20 participants, resulting in a total of over 100 participants. The number of participants was limited and the workshops were not provided in a hybrid manner, as their nature is highly interactive. Participants had varying levels of familiarity with CLARIN.SI and CLASSLA resources. After the workshop, the participants were surveyed regarding their intentions to use the CLASSLA-web corpora in their research endeavors. They intend to use the corpora for various research purposes, including collocation analysis, frequency analysis, lexicography, comparisons between languages, specialized language studies (pertaining to child language, loanwords, phraseology, gendered language and vulgarisms), as well as for natural language processing tasks. Additionally, some plan to incorporate corpora into their academic work, including teaching.

Date	**Location**	**Venue**
19/4/2024	Zagreb, Croatia	Faculty of Humanities and Social Sciences, University of Zagreb
26/4/2024	Rijeka, Croatia	Center for Language Research, Faculty of Humanities and Social Sciences, University of Rijeka
29/5/2024	Belgrade, Serbia	International conference Leksikografski susreti
4/6/2024	Skopje, North Macedonia	Blaže Koneski Faculty of Philology, Ss. Cyril and Methodius University
26/6/2024	Sofia, Bulgaria	International CLaDA-BG Conference 2024
18/9/2024	Ljubljana, Slovenia	Language Technologies & Digital Humanities Conference 2024

Table 1: Details of the first iteration of the CLASSLA-Express workshops.

4 Promotional Activities

An important part of the organization is dissemination of information about the workshops. Firstly, a dedicated web page[7] was created and published within the CLARIN.SI website to centralize information about the workshops. The dissemination of the calls for participation involved both traditional academic channels, such as posting news items on various local and international mailing lists, as well as social media platforms, specifically the X[8] and LinkedIn[9] profiles of CLARIN.SI. Leveraging social media enabled a wider outreach to individuals across diverse fields, expanding the dissemination beyond the linguistic community. A survey distributed among the participants after each workshop showed that most of them learned about the CLASSLA-Express workshop via mailing lists, however, many participants found out about the workshop by word of mouth or on social media.

After the workshops, we shared the key highlights from each event on the CLARIN.SI website[10] and on social media platforms.

Moreover, efforts were made to enhance the workshops' visibility. We designed a logo for the workshop, shown in Figure 1. As an additional touch, certificates of attendance and logo-branded bags were prepared and distributed to participants.

5 Extending the Workshop Series to New Venues

The first iteration of the CLASSLA-Express workshop series was mainly convened by the two lecturers who were willing to travel and conduct the workshop in five countries. However, the workshops can be easily extended to new locations in collaboration with colleagues at those venues, as was done with the organizer in Sofia, Bulgaria, a stop added after the first five stops were already agreed upon. We are prepared to offer assistance by providing templates for calls for participation, registration forms and

[7] https://www.clarin.si/info/k-centre/workshops/classla-express/

[8] https://twitter.com/ClarinSlovenia

[9] https://www.linkedin.com/company/clarin-si

[10] https://www.clarin.si/info/k-centre/workshops/

Figure 1: The CLASSLA-Express logo

certificates of attendance, as well as by promoting the workshops on the CLASSLA-Express website, and CLARIN.SI and CLASSLA communication platforms. Teaching materials have been developed in Croatian, Bulgarian and English and are available for sharing with interested partners. Given that CLASSLA-web corpora cover all South Slavic languages and have comparable format and scope, the queries designed for Slovenian, Croatian, Serbian, Macedonian and Bulgarian can be easily adapted to other South Slavic languages and applied to a specific CLASSLA-web corpus.

Following the announcement of the calls for participation for the initial CLASSLA-Express workshops, colleagues from Montenegro, and Bosnia and Herzegovina, as well as from additional locations in Croatia and Serbia, expressed interest in hosting workshops in their respective countries and cities, and these additional workshops are already in sight.

6 Conclusion

In this paper, we presented the strategy employed by the CLASSLA-Express workshop series to promote the CLARIN.SI linguistic resources and infrastructure, as well as the CLASSLA Knowledge Centre, beyond national borders. The workshops' approach involves directly engaging countries with interested audiences, eliminating the need for individuals to travel to a specific location. Additionally, in many locations, the workshops were conducted in the national languages. This method not only facilitates accessible and environmentally-friendly knowledge sharing but also broadens the reach of diverse individuals from different backgrounds and nations who become familiar with the resources and infrastructure. Furthermore, the dissemination of workshop information through various channels, including social media platforms, aims to engage individuals beyond the existing CLARIN.SI academic community, thereby expanding awareness of the resources and activities to a wider audience. Secondly, the workshop series is designed in such way that it can be easily extended to other locations. The CLASSLA-web corpora collection (Ljubešić & Kuzman, 2024) which is the main focus of the workshops, comprises comparable corpora for all South Slavic languages. This allows the practical tasks to be adapted to any South Slavic CLASSLA-web corpus. In addition, we offer teaching materials in multiple languages to interested collaborators who wish to host the workshop in their own city. We also assist them with the logistics by providing templates for calls for participation, certificates of attendance, registration forms, and promoting the event through various channels.

The train of workshops has already gained momentum, with additional workshops in Montenegro, Bosnia and Herzegovina, Croatia and Serbia in sight. In the future, we plan to further extend the scope of the workshops by covering approaches on how to support corpus-based research with large language models, a discussion topic reappearing in each of the conducted workshops.

Acknowledgments

The CLASSLA-Express workshop series was funded by the CLARIN.SI infrastructure. The work in preparing and supporting the workshop series was partially funded by the programme P6-0411 "Language Resources and Technologies for Slovene" financed by the Slovenian Research and Innovation Agency (ARIS). The workshop in Sofia was also partially supported by CLaDA-BG, funded by the Ministry of Education and Science of Bulgaria.

References

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 29–34. https://doi.org/10.18653/v1/W19-3704

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

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024a). Bulgarian web corpus CLASSLA-web.bg 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1928

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024b). Croatian web corpus CLASSLA-web.hr 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1929

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

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024d). Serbian web corpus CLASSLA-web.sr 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1931

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024e). Slovenian web corpus CLASSLA-web.sl 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1882

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *arXiv preprint arXiv:2308.04255*.

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

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

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

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

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

Abstract

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

1 Introduction

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

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

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

the landscaping activity of Work Package 2. This paper aims to show the application of the training strategy detailed in WP8 to the modules designed by CLARIN-IT within activity 8.2, “Teach CLARIN, teach with CLARIN: training, communication and impact” adapting them to the training needs of the University of Ferrara. This experience will result in a case study to show the potential of such a strategy when integrating the CLARIN research infrastructure into teaching.

2 Contextual Background

2.1 H2IOSC Training Methodology Applied to CLARIN Courses

H2IOSC's training strategy is aimed primarily at the Italian Social Sciences and Humanities communities and offers tailored training on the resources and tools available in the participating RIs. An essential pillar of this strategy is the train-the-trainers perspective to help teachers of university and professional courses enhance their skills and those of their students. In the first half of the H2IOSC project, the CLARIN-IT consortium, represented by its founder member and host of the ILC4CLARIN service provider centre at the CNR Institute of Computational Linguistics[3], worked on applying the shared H2IOSC methodology for the design of training materials as FAIR digital objects[4] to the UPSKILLS course “Introduction to Language Data: Standards and Repositories” which was already reusable in compliance with the FAIR principles and accessible on the project’s Moodle platform[5]. Starting from the second trimester of 2024, H2IOSC Activity 8.2 promoted training initiatives among the members already part of the CLARIN-IT consortium. The first initiatives included a training module tailored to the requests of the ITSERR project’s research team working on the consolidation of the ESFRI RI RESILIENCE[6] and the training module directed to Dr. Dario del Fante, researcher and teacher at the University of Ferrara. Both these events were held online, taught by the dedicated H2IOSC personnel and resulted in further collaboration on specific requests, which, in the case of Ferrara, included the repurposing of selected lessons from the UPSKILLS course “Introduction to Language Data: Standards and Repositories” for three-degree programs which will be shown in Section 3.

2.2 Teaching Activity at the University of Ferrara

Three English Language and Linguistics courses have been identified as pedagogical spaces to integrate CLARIN resources and services into the teaching material. Each course is part of three different degree programs held at the University of Ferrara: 1) Bachelor's degree in Humanities, Arts, and Archeology; 2) Master degree in Foreign Languages and Literatures; 3) Bachelor's degree program in Manager of Cultural Itineraries.

The English Language for Humanities, Arts, and Archeology course aims to provide students with critical tools to understand English information. The course focuses on three main topics: English critically approached as a global language, the notion of discourse, and communicative strategies for representing people and identities in political communication and social media contexts. The English Language course for Modern Languages and Literature consists of a theoretical part, which offers a panorama on the main aspects of Cultural Translation, and a practice section, which will translate literary texts of various periods and contexts from English to Italian. The English Language for Tourism course aims to equip students with linguistic and metalinguistic tools essential for developing an awareness of the English tourist language. This goal is achieved through linguistic and cultural analysis of various contemporary tourism text types.

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

[4] The H2IOSC shared methodology for design and adaptation of FAIR training materials is detailed in Pedonese, Frontini et al., 2024 (*in press*).

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

[6] https://www.itserr.it/. RESILIENCE is the European RI for religious studies (https://www.resilience-ri.eu/)

Integrating CLARIN materials and services is based on recognising that the world has evolved into an increasingly digitised context. Within the humanities and social sciences, every role must recognise the engagement with the digital. In this way, the decision to implement CLARIN materials is linked, on the one hand, to guiding students towards digital thinking and familiarising them with the digital world. On the other hand, the aim is to provide tools and resources that can concretely aid them in their work. To this end, we included two lessons at the end of each course after covering the most important topics and providing a brief introductory approach to the world of the CLARIN infrastructure.

3 CLARIN-IT Training at the University of Ferrara

The training initiative stemmed from pre-existing contacts between the CNR-ILC and the University of Ferrara, a CLARIN-IT consortium member — and was detailed following the steps of the training management system developed within the H2IOSC project: the applicants were asked to fill in a form to provide organisational details such as the preferred venue of the training, date, duration, general topics of interests along with specific requests on hands-on sessions on the CLARIN services. Among the topics of interest were FAIR principles, CLARIN core services and translation technologies for teaching and a first meeting online was scheduled to update Dr. Dario del Fante, as part of the University of Ferrara team, on the possible applications of CLARIN services in the courses in which he was teaching (Section 1.2). The following subsections will detail the trainer's profile and the activities carried out in each course.

3.1 Training the Trainer

The first meeting was held online and consisted of the presentation of the primary teaching resources currently available for the Italian SSH community. The materials from the adapted UPSKILLS course and the facilitation guide developed within the H2IOSC project on how to reuse them were presented. Another precious resource in this phase was the UPSKILLS *Guidelines on integrating Research Infrastructures into Teaching* (van der Lek et al. 2023), an up-to-date tool to enhance research-based teaching, presenting the case study of CLARIN. Since Dr. Del Fante was already familiar with CLARIN core services, the training session focused mainly on their latest functionalities and on the discovery process of language resources to teach students translation courses. Before scheduling another meeting, the parties agreed to collaborate to include H2IOSC personnel in creating a joint module to be held at the University in May.

Upon requests of specific translation technologies to teach students, further inquiries were made by the CLARIN-IT trainer, which resulted in a fruitful discussion with the CLARIN ERIC training office and the outline of a CLARIN Café on Translation Technologies and Workflows for SSH Research held on June 14, 2024, and aimed at providing instructors, researchers and CLARIN infrastructure staff the tools on how to introduce these technologies in their classrooms and research projects[7].

3.2 Training the Students

We developed three modules consisting of two lessons each to be held in person in Ferrara within the three courses mentioned above: 1) English Language Course for Tourism: two lessons of two hours each for 40 students; 2) English Language for Translation: two lessons of two hours each for 25 students; 3) English Language and Linguistics for Humanities, Arts and Archaeology: two lessons of two hours each for 30 students.

For the first lesson of each module, we adopted the same format, tailoring the examples and case studies to the disciplinary requirements of each course. This introductory lesson offered a detailed overview of the CLARIN research infrastructure, emphasising the support provided by CLARIN Knowledge Infrastructure and providing information on funding opportunities. Additionally, we introduced the use of CLARIN Language Resource Families and the main functions of the metacatalogue Virtual Language Observatory (VLO). We then dedicated the second lesson of each module to the practical application of CLARIN core services, and, in this case, we tailored the content

[7] https://www.h2iosc.cnr.it/h2iosc-training-clarin-cafe/

to the specific needs of each course. Each resource and tool we chose to show the students was accessed through the Virtual Language Observatory to familiarise the class with CLARIN services, and we encourage students of all three courses to include those digital tools in their research projects.

In the first course, we implemented the recently released Parliament 4.0 to analyse cross-linguistic political communication in Italy and the UK: we demonstrated how to use metadata to analyse textual data differently and provide a more comprehensive analysis of the texts. In the second course, we focused on using the parallel corpora available in the CLARIN infrastructure to assist translators. Specifically, we utilised the Intercorp parallel corpus (Čermák & Rosen 2012) to provide exercises addressing translation issues. We also utilised *Treq - the translation equivalent database*[8] - to compare the best translation equivalents and synonyms available in the corpus. For the third course, we demonstrated the utility of Voyant Tools in analysing the most recurrent features of promotional language used on different international tourist websites.

4 Results and Future Developments

This series of training events at the University of Ferrara allowed us to collect insight into using CLARIN language resources and services in academic teaching activities and how to develop new modules re-using existing courses, namely the UPSKILLS training materials. Thanks to this experience, we could better understand the steps needed to train a trainer who is already familiar with CLARIN resources but has new and challenging requests to which CLARIN-IT needs to respond with the help of the broader international consortium. Another lesson learnt is the necessity of adapting methods, standards, tools and resources to make them relevant to the Italian community, which is only possible thanks to disseminating success stories on both national and international levels. Regarding student feedback, we prepared a final questionnaire with evaluation questions on the knowledge acquired during the course and questions on quality satisfaction. We received a total of 19 responses. As for the knowledge assessment, most students answered the questions correctly, showing they had learnt about topics such as Open Science principles, licence use, and CLARIN core services. The average student satisfaction with the course was 3.63 out of 5, indicating that the course was calibrated to the students' abilities, who performed well in the knowledge assessment. The only point where improvement should be made concerns the students' perception of the relevance of that knowledge since the course collected an average score of 3.42 on usefulness. In this respect, we could better explore the application of CLARIN tools and services from the perspective of future professionals employed in fields such as translation, publishing or the cultural industries, by documenting the need for data stewardship outside the field of research, so as to better tune the course to specific needs of certain professional categories. Due to the lack of space in this abstract, we will better show the survey results in the presentation.

References

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

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

'FAIR Principles'. n.d. GO FAIR. Accessed 12 January 2024. https://www.go-fair.org/fair-principles/.

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

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

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

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

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

'Integrating Research Infrastructures into Teaching: Recommendations and Best Practices'. n.d. Accessed 29 November 2023. https://doi.org/10.5281/zenodo.8114407.

Pedonese Giulia, Frontini Francesca, Ottaviani Roberta, Boschetti Federico, Spadi Alessia, Francalanci Lucia, Scognamiglio Alessia, Restaneo Pietro, Chaban Antonina, Striova Jana, Benassi Laura 'Materiali didattici come oggetti digitali FAIR: una metodologia condivisa per la formazione in H2IOSC' (AIUCD 2024), pp. 577-581, *in press*.

'Recommendations for a Minimal Metadata Set to Aid Harmonised Discovery of Learning Resources'. 2022. RDA. 1 April 2022. https://www.rd-alliance.org/group/education-and-training-handling-research-data-ig/outcomes/recommendations-minimal-metadata-set.

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

Federated Content Search: Advancing the Common Search Infrastructure

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

Abstract

About twelve years ago, the idea and general architecture of a "federated search" as part of the CLARIN research data infrastructure was presented for the first time. Since then, the Federated Content Search (FCS) has continuously come closer to its original goals of a powerful search infrastructure based on large amounts of distributed research data. In the last two years in particular, development has accelerated massively, new application scenarios have been formulated, new user groups have been opened up, new tools and user interfaces have been developed. This paper assesses to which degree the original goals have been achieved, provides a broad overview of the exciting developments of recent years and the topics that are currently being worked on.

1 Introduction

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

In addition, various topics were described as possible future priorities, including the provision of bindings for popular data indexing systems, integration with other CLARIN services, support for access-restricted resources via suitable control mechanisms, and the use of ISOcat (Kemps-Snijders et al., 2009) as a semantic integration layer between the common infrastructure and specific endpoint implementations. Version 1.0 of the FCS specification (Schonefeld et al., 2014) two years after the original paper incorporated large parts of these ideas and was superseded by version 2.0 (van Uytvanck et al., 2017) in 2017 that introduced a new query language and a new data view to enhance access to annotated text corpora.

In the last years, the general development of the overall FCS ecosystem has changed significantly. This is characterised, among other things, by the shift of development processes to popular working environments (GitHub/GitLab[1]) for supporting modern and open development processes, improved documentation based on easily editable formats (e.g. using AsciiDoc), and an increasing number of software libraries for a wide variety of use cases and technology stacks.

The FCS continues to be a central component of the CLARIN infrastructure and is a prominent part of its current work plan, but has also established itself in institutions and projects beyond this. A prominent

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

example is the establishment of the FCS as the central search component in the German infrastructure consortium *Text+*[2] (Hinrichs et al., 2022), which has already significantly increased the number of available resources. In the same context, new user interfaces are currently being developed to improve usability and user-friendliness. In addition, new usage scenarios are increasingly being supported that benefit from the flexible extensibility of the FCS standard.

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

2 Extensions

Extensibility is a key factor and an important design principle for the FCS. As a consequence open standards were deliberately chosen to the largest degree possible. This allows the FCS to adapt to evolving user requirements while still keeping backwards compatibility for (older) endpoints and clients. Multiple extension proposals are currently being worked on, including the integration of additional resource types and managing access to restricted data. Figure 1 summarises the overall FCS architecture, highlighting extensions of its ecosystem that are currently in progress.

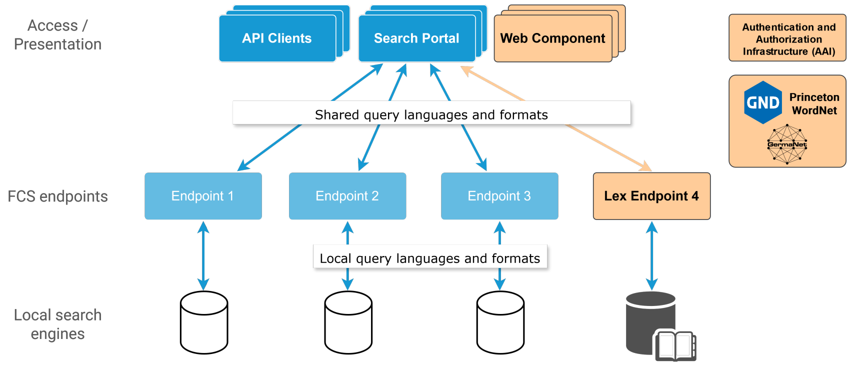


Figure 1: General architecture of the FCS, highlighting currently developed extensions in yellow.

2.1 LexFCS

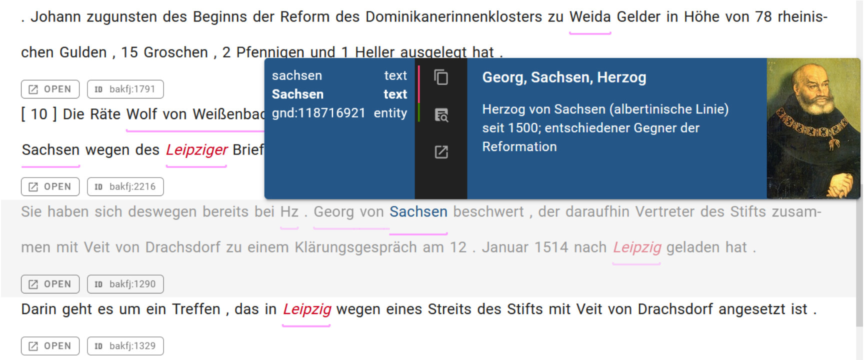
The FCS was originally intended for searching primarily in full-texts with optional annotation layers. This, however, excludes differently structured resource types such as lexical resources, including dictionaries, word lists or semantic wordnets. A dedicated working group in Text+ prepared a working proposal for the new *LexFCS* extension (Eckart et al., 2023). The alpha specification (Körner et al., 2023) proposes a new query language for key-value structured resources and a slight adaption of the standard result format allowing optional semantic annotation. Current work is focusing on an advanced lexical data view to allow improved presentation of complex records and the standardisation of terminology for queries and results. The Text+ FCS Aggregator[3] already includes a first implementation of this extension.

2.2 EntityFCS

Authority files and wordnets are important knowledge bases when annotating linguistic resources. The *entity-oriented search* paradigm (Balog, 2018) focuses on accessing research data on their basis. The *EntityFCS* extension proposes a combined search on entity- and sense-annotated full texts as well as lexical resources using global identifiers from authority files like GND (German National Library (DNB), 2024), WikiData (Vrandečić & Krötzsch, 2014), GermaNet (Hamp & Feldweg, 1997), or Princeton

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

WordNet (Fellbaum, 1998). Entities may include persons, organisations, events, and locations among other types. This allows the utilisation of already annotated data and enables a more generic search to e.g. cover alternative written forms beyond resource type or language boundaries (including historical variants) or the disambiguation of homonyms.



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

OPEN bakfj:1791

[10] Die Räte Wolf von Weißenbac

Sachsen wegen des *Leipziger* Brief

OPEN bakfj:2216

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

Georg, Sachsen, Herzog

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

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

OPEN bakfj:1290

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

OPEN bakfj:1329

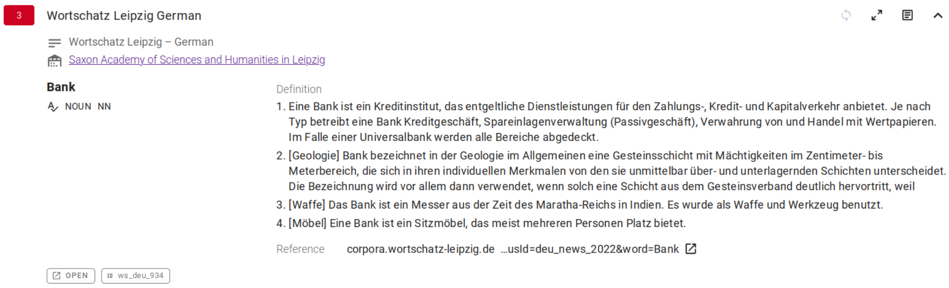
Figure 2: Presentation of complex annotations using visual highlights and descriptive tooltips for entities, loaded via an external API,

2.3 Further FCS Extension Proposals and Ongoing Work

More FCS extensions are currently being worked on, but are still in the user requirements analysis phase or only consist of initial prototypes for evaluation and testing purposes.

One frequently requested feature is the support of access-restricted resources via the established Authentication & Authorization Infrastructure (AAI) which is a long-standing requirement with an early specification draft and a first implementation. Further user requests for reference-only results or support of *derived text formats* (Schöch et al., 2020) are being discussed for cases where the actual content cannot be provided due to contractual or legal reasons. The potential support for queries on syntactic structures (e.g., in dependency treebanks) is a fairly recent development. A proposal for such a new type of query language might lead to a new *SyntacticFCS* extensions.

Another open aspect is the improved description of results via structured metadata. The FCS specification is currently very limited on how much additional metadata is available and can be presented to users. Data providers with resources that, for example, aggregate texts from various sources such as newspaper collections, require methods to enrich individual results with more descriptive metadata, including publication and release date, title, author information and more.



3 Wortschatz Leipzig German

Wortschatz Leipzig – German

Saxon Academy of Sciences and Humanities in Leipzig

Bank

NOUN NN

Definition

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

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

OPEN ws_deu_934

Figure 3: Representing the micro structure of a dictionary entry using the new lexical data view (user interface prototype)

3 Usability, Software Ecosystem, End-user & Developer Support

A vibrant ecosystem requires a variety of factors to support its continuous development and a growing amount of resources and new users. This includes user-friendly interfaces, end-user support for scientists using the FCS, support for developers working on new endpoints and central components, suitable communication platforms and documentation formats, and much more. A large number of these activities have been carried out in recent years, which are briefly described below.

Improved Users Interfaces. The *FCS Aggregator* has been revised several times since its first release and has been continuously developed further. The latest major changes include the improved provision of a powerful RESTful API so that its functionality can also be used in other applications. On this basis, there is now a Web component developed using the `Vue.js` framework that enables other forms of access to FCS endpoints and the presentation of results via an alternative Web interface. These include a first integration of the EntityFCS based on the GND (cf. Figure 2), the improved presentation of lexical resources (cf. Figure 3), but also the option of integrating this new search interface into your own application with little effort. This is already implemented as part of the FCS integration in the central Text+ portal[4] and in first repository websites that want to support searching in their local resources[5].

End-user & Developer Support. Various channels and activities are offered to support end-users, endpoint developers and operators. This includes the CLARIN Forum[6] to transparently publish news as well as allowing interaction and user feedback, workshops for FCS endpoint development (e.g., organised by the Saxon Academy of Sciences and Humanities in Leipzig (SAW) for Text+[7]) and help desks. Comparable hackathons for developers and workshops for end users focusing on specific usage scenarios are being actively planned. All information material created in the process will be made available on the respective channels as well.

Documentation. Code and documentation was moved to GitHub.com to make it more accessible. The working format for the FCS specification and other documents[8] is now the AsciiDoc format that supports features such as improved editing, cross-linking and various output formats. Additional material like FCS Endpoint Development tutorials and extensive presentation slides[9] – focusing on different user groups – were created to ease development efforts.

Libraries and Software Ecosystem. Reference libraries were initially developed in `Java` and have now also been translated to `Python` to open up the ecosystem for FCS endpoints and applications. They are being continuously improved and extended. Numerous endpoint implementations in various other languages exist, with open-source ones listed in the *Awesome FCS* list[10].

Tools and Validators. To support FCS users and developers, various tools are provided for development, with many of them *dockerized* to ease setup. An important application is the **FCS Validator**[11] that has been completely rewritten and extended to cover more test cases. It now offers additional features for configuration and reporting in response to evolving user requirements.

Acknowledgments

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

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

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

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

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

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

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

[10] A curated list of FCS frameworks, libraries, software and resources, https://github.com/clarin-eric/awesome-fcs

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

References

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

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

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

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

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

Hinrichs, E., Trippel, T., Hinrichs, M., Illig, E. M., & Witt, A. (2022, January). CLARIN Café on Text+: A New Research Data Initiative in Germany. https://www.clarin.eu/event/2022/clarin-cafe-research-data-infrastructure-text

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

Körner, E., Eckart, T., Herold, A., Wiegand, F., Michaelis, F., Bremm, M., Cotgrove, L., Trippel, T., & Rau, F. (2023, May). *Federated Content Search for Lexical Resources (LexFCS): Specification*. https://doi.org/10.5281/zenodo.7849753

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

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

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

Stehouwer, H., Durco, M., Auer, E., & Broeder, D. (2012). Federated Search: Towards a Common Search Infrastructure. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3255–3259. http://www.lrec-conf.org/proceedings/lrec2012/pdf/524_Paper.pdf

van Uytvanck, D., Olsson, L.-J., Schonefeld, O., Eckart, T., Körner, E., Kisler, T., Fischer, P. M., & Illig, E. M. (2017). *CLARIN Federated Content Search (CLARIN-FCS) – Core 2.0*. https://office.clarin.eu/v/CE-2017-1046-FCS-Specification-v20230426.pdf

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

The citation of language resource technologies in CLARIN

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

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

Abstract

This paper first presents a use case of citation practices of the CLARIN community from the perspective of the *Joint Declaration of Data Citation Principles*. The paper subsequently reviews the existing data citation guidelines of CLARIN B-centres, focusing on authorship attribution as well as other principles such as access and persistence. On the basis of the review, the paper puts forward an explicit proposal that seeks to account for identified gaps in both citation practice and existing guidelines.

1 Introduction

This paper presents and discusses how the CLARIN community typically cites language resource technologies (LRTs), such as corpora, lexical datasets, and software. We focus on how the citations adhere to the *Joint Declaration of Data Citation Principles* (Martone, 2014), specifically to guidelines concerning Authorship Attribution, as well as other principles such as Access and Persistence. We also review the guidelines offered by CLARIN B-centres for data citation. On the basis of the identified inconsistencies in citation practice and the gaps in instructions, we draw up an explicit proposal for LRT citation.

The paper is structured as follows. In Section 2, we present common CLARIN citation practices by using the extended abstracts published in last year's proceedings of the annual CLARIN conference (Lindén et al., 2023a) as a use case. In Section 3, we review the citation guidelines of those CLARIN repositories that currently have B-centre certification. Section 4 ends the paper with the proposal.

2 Citing LRTs in practice

The results of the survey of citing LRTs in the abstracts published in (Lindén et al., 2023a) are given in Table 1. What we have looked at is how any kind of LRT, be it a language corpus, lexical resource, tool, or language model, is cited in the abstracts. The Table shows that there are overall 189 citations across the 37 extended abstracts published in the proceedings, while citations are grouped together under 5 major citing strategies.[1]

Citation strategy	#	%
LRT citation	20	11%
Paper about LRT	74	39%
Paper & LRT citation	7	4%
URL in text	76	40%
Paper & URL in text	12	6%
Σ	189	100%

Table 1: Citing practices in *CLARIN Conference Proceedings 2023* (Lindén et al., 2023a)

The first strategy is what we label 'LRT citation', which makes up 20 (or 11%) of the citation cases. This refers to those cases where the actual LRT is cited, with full authorship attribution and a URL to the repository where the LRT or its metadata are available. Several examples of this strategy are found in the paper by Xu et al. (2023), who for instance cite the *Icelandic pronunciation dictionary for language technology* by providing the following reference: (Nikulásdóttir, 2021). Because this sort of citation ensures the proper findability, identifiability and authorship attribution of the LRTs, it is – along with the

[1] See the full results in this Google Spreadsheet.

‘Paper & LRT citation’ strategy – the most in line with the *Joint Declaration of Data Citation Principles* (Martone, 2014).

The second strategy is labelled ‘Paper about LRT’, which is the second most common one (74 or 39% of all cases). An example of this is found in the paper by Janssen (2023), who cites the tool *UDpipe* by providing a reference to the paper about the tool, which is (Straka & Straková, 2017). What the author does not do, however, is provide a reference for the tool itself, which is (Straka & Straková, 2016).

The third strategy is a mix between the first two – that is, citing both the LRT as well as the paper about it, which in the case of *UDpipe* would mean providing a reference both to the paper (Straka & Straková, 2017) and the tool (Straka & Straková, 2016). This is the least common strategy encountered.

The fourth strategy is simply providing a URL to the tool somewhere in the text, most commonly as a footnote. Note that persistent identifiers (PIDs) in this case make up a minority of the URLs provided. An example of this is found in the paper by Lindén et al. (2023b), who provide in a footnote the PID http://urn.fi/urn:nbn:fi:lb-2022062221 for their corpus *FINDarC*, but do not provide a reference for the corpus itself, which would be (Harviainen, n.d.) according to the instructions in the FIN-CLARIN repository.

The last strategy is again a mix, this time of providing a URL as well as the paper about the tool.

In sum, the citation of LRTs by the CLARIN community appears to be quite varied in practice, with ‘proper’ LRT citations being in the minority. For this reason, we have also surveyed what kind of citation instructions are provided across the CLARIN infrastructure. We turn to this survey in the next section.

3 Citation instructions of CLARIN B-centres

Centre	LRT Citation	Author	Year	Publ.	PID	BibTeX
CLARIN-IS	✓	✓	✓	✓	✓	✓
CLARIN-LV	✓	✓	✓	✓	✓	✓
CLARIN-DK	✓	✓	✓	✓	✓	✓
CLARIN-PL	✓	✓	✓	✓	✓	✓
CLARINO	✓	✓	✓	✓	✓	✓
CLARIN.SI	✓	✓	✓	✓	✓	✓
LINDAT	✓	✓	✓	✓	✓	✓
ILC4CLARIN	✓	✓	✓	✓	✓	✓
ACDH-CH	✓	✓	✓	✓	✓	✓
FIN-CLARIN	✓	?	?	✓	✓	✓
ORTOLANG	✓	?	✓	✓	✓	✓
MPI	✓	?	✓	✓	✓	×
CLARIN:EL	✓	?	✓	✓	✓	×
CLARIN-D/ASV	✓	×	?	✓	✓	✓
CLARIN-BE	✓	×	✓	✓	✓	×
CMU-TalkBank	×	✓	×	×	✓	NA
ZIM	×	✓	✓	✓	✓	NA
CLARIN-D/UdS	×	✓	✓	?	✓	NA
CLARIN-D/BAS	×	?	✓	✓	✓	NA
PORTULAN	×	?	✓	×	✓	NA
CLARIN-D/BBAW	×	×	×	×	×	NA
CLARIN-D/SFS	×	×	×	×	×	NA

Table 2: Citation guidelines across CLARIN B-centres

Table 2 provides the survey of the citation instructions of all the 22 repositories that are as of 22 April 2024 certified as CLARIN B-centres. The first thing that we have checked is whether a centre provides a dedicated citation box for the deposited LRT. It turns out that most of the B-centres do so (see the ‘LRT citation’ column), with the 7 centres listed below the dashed line in the Table being the exception. An interesting case here is CMU-TalkBank, since the webpages of the corpora available there do include citation instructions, but they do not instruct researchers to cite the corpora themselves, but rather one of the papers associated with them; for instance, for the *Dresden Corpus* (see Kubanek-German, n.d.), which is a L2-learner corpus, the instructions are just to cite the following paper: (Kubanek-German, 2000), which is not even a paper about the corpus.

Additionally, the Table shows the inclusion of 4 types of metadata relevant for citations – human authorship, year of publication, publisher (i.e., the name of the repository), and the persistent identifier (PID) – as well as whether the centre provides a BibTeX export for the citation. Out of these 4 metadata categories, the Author field stands out the most, since this information is sometimes fully missing (the centres marked by ×) or the centres are inconsistent with what they consider to be authorship – i.e., instead of listing the people who have worked on the corpus, they often include non-human entities as authors (centres marked with ?). An example of the latter is found in the citation instructions for *The CLES corpus of spontaneous L2 English* (LIG et al., 2024) in ORTOLANG, where LIG in the citation stands for 'Laboratoire d'informatique de Grenoble'; indeed, all entities listed as authors in the metadata of this corpus are non-human entities such as laboratories and institutes.

The other 3 citation metadata – year of publication, publisher, and the PID – are more readily included; year of publication is missing only in the case of 3 centres (CMU-TalkBank, CLARIN-D/BBAW, and CLARIN-D/SFS); explicit instructions for citing the publisher – that is, the repository – are missing in the case of 4 centres (CMU-TalkBank, PORTULAN, CLARIN-D/BBAW, and CLARIN-D/SFS), and only 2 centres do not provide PIDs for their LRTs (CLARIN-D/BBAW; CLARIN-D/SFS).

4 Proposal

To ensure compliance with the *Joint Declaration of Data Citation Principles* (Martone, 2014), we propose that the rule in (1) should be followed by anyone who cites any kind of LRT.

(1) **Citing LRTs – proposal**
Cite any kind of language-resource technology, be it a language corpus or lexical resource, or any kind of tool, software or language model, as you otherwise would a book, but make sure to provide the PID as well.

What is meant by the analogy with books in rule (1) is twofold. First, we believe that the necessary bibliographic elements when citing an LRT are, just like with books, only the following four: the title, human authorship, the publisher – that is, the repository where the LRT is deposited –, and the year of publication. The only addition that we propose for citing LRTs is the obligatory inclusion of the PID, which is otherwise typically not provided in the case of book citations. It is important to note that most of the CLARIN centres that use the CLARIN DSpace infrastructure (Straňák et al., 2020), so all the centres from CLARIN-IS to ILC4CLARIN in Table 2, along with ACDH-CH, already provide such instructions. The centres that do not are primarily those marked with ? or × under the Author column, since they either inconsistently provide instructions for citing human authorship for their deposits or do not do so at all. But it is human authorship rather than a list of departments, institutes, or laboratories that most precisely adheres to the Credit & Attribution section of the *Joint Declaration of Data Citation Principles*, specifically to the idea that 'data citation should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data' (Martone, 2014).

What is interesting to observe in this respect is that there is some variation in additional types of citation metadata provided by the centres. For instance, ORTOLANG also provides – as a separate citation element – the version of the LRT, while in the case of e.g. CLARIN.SI, the version is simply given as part of the name of the LRT, as is for instance the case of *ParlaMint 4.0* (Erjavec et al., 2023). ORTOLANG also separately specifies in the citation instruction the type of the LRT – that is, the difference between *corpus*, *dataset*, *tool*, and so on. For citing the LRT version, we believe that the easiest solution is simply to adopt the CLARIN.SI (or broader CLARIN DSpace) strategy and list it as part of the LRT name. Note that the issue of labelling different versions becomes simply a matter of convention (rather than a necessary condition for identification) if each new version of an LRT is assigned a unique PID. Indeed, such an approach to versioning is already an established practice in CLARIN.[2] Similarly, LRT type can also be (and in fact often is) specified as part of the name. In terms of related work, we thus advocate for the usage of Conzett and De Smedt's (2022) **minimal bibliographic template** (i.e., *author, date, title, publisher,* and *locator* being sufficient citation elements, with the proviso that authors should correspond to

[2] See *ParlaMint 4.0* (Erjavec et al., 2023) vs. *ParlaMint 4.1* (Erjavec et al., 2024) for an example.

human individuals), but not for the extended one, which also specifies three additional citation elements (*Other Attribution (Roles)*, *Version*, *Date Accessed*).[3]

The second point of the analogy with books has to do with the stylistic formatting of the bibliographic entry. A characteristic feature of formatting a book entry is the italicised title, a convention which remains constant between different citation standards. Interestingly, the stylistic formatting of LRTs as books is already implicit in for instance the CLARIN DSpace repositories, where the convention of typesetting the title is to italicise it; see for instance the citation instructions for *ParlaMint 4.0* in the yellow box in (Erjavec et al., 2023). There is, however, a discrepancy with all the BibTeX exports in Table 2, as they invariably define all citations as a **Misc** entry rather than a **Book** entry. In the case of APA, for instance, **Misc** entries lose the italic formatting of the title, and quite generally do not conform to any well-defined bibliographic type. Note, however, that BibTex also defines a **Dataset** type, which importantly formats the bibliographic entry in the exact same way as a **Book** entry. If (1) is adopted, a BibTex export could then be as in (2), defined for *ParlaMint 4.0* (with the long list of authors abbreviated).

(2)
```
@dataset{11356/1859,
title = {Multilingual comparable corpora of parliamentary debates {ParlaMint} 4.0},
publisher = {Slovenian language resource repository {CLARIN}.{SI}},
author = {Tomaž Erjavec and others},
year = {2023},
url = {http://hdl.handle.net/11356/1859}
```

Aside from substituting **Misc** with **Dataset** (another option would simply be to use **Book**, given that **Book** and **Dataset** are equivalent in BibTex), the repository is now defined as the publisher rather than under the Note field. Importantly, the book (or BibTex's **Dataset/Book**) format is well defined across all citation standards, so this would also help with the possible creation of a widget to export the instructions automatically into the different standards, which is currently not possible in CLARIN repositories.

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)*, the DARIAH-SI Infrastructure (Slovenian Research Agency project *I0-E007*), as well as the CLARIN Resource Families Initiative.[4]

References

Conzett, P., & De Smedt, K. (2022). Guidance for citing linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 143–155). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0015

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., ... Fišer, D. (2023). *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1859

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., ... Fišer, D. (2024). *Multilingual comparable corpora of parliamentary debates ParlaMint 4.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1912

Harviainen, T. (n.d.). *Finnish dark web marketplace corpus*. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2022062221

[3] *Date Accessed* seems redundant in the case of PIDs, as the latter are long-lasting by nature. *Other Attribution (Roles)* refers to acknowledging contributors other than authors and is according to Conzett and De Smedt (2022) subfield specific (e.g., acknowledging language consultants in language documentation).

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

Janssen, M. (2023). Dynamically chaining APIs: From Dracor to TEITOK. *CLARIN Annual Conference Proceedings*, 116–119.

Kubanek-German, A. (n.d.). *Dresden corpus*. TalkBank. https://hdl.handle.net/10.21415/T52G75

Kubanek-German, A. (2000). Early language programmes in Germany. In *An early start: Young learners and modern languages in Europe and beyond* (pp. 59–70). Council of Europe Publishing Strasbourg.

LIG, LIDILEM, ILCEA4, & STL. (2024). *The CLES corpus of spontaneous L2 English*. ORTOLANG. https://hdl.handle.net/11403/cles-spontaneous-english/v1

Lindén, K., Niemi, J., & Kontino, T. (Eds.). (2023a). *CLARIN annual conference proceedings*. https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf

Lindén, K., Ruokolainen, T., Hämäläinen, L., & Harviainen, J. T. (2023b). Sharing the Finnish dark web marketplace corpus (FINDarC). *CLARIN Annual Conference Proceedings 2023*, 134–139.

Martone, M. (Ed.). (2014). *Data citation synthesis group: Joint declaration of data citation principles*. FORCE11. https://doi.org/10.25490/a97f-egyk

Nikulásdóttir, A. B. (2021). *Icelandic pronunciation dictionary for language technology*. CLARIN-IS. http://hdl.handle.net/20.500.12537/99

Straka, M., & Straková, J. (2016). *UDPipe*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal; Applied Linguistics (ÚFAL), Faculty of Mathematics; Physics, Charles University. http://hdl.handle.net/11234/1-1702

Straka, M., & Straková, J. (2017). Tokenizing, PoS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 shared task*, 88–99.

Straňák, P., Košarko, O., & Mišutka, J. (2020). CLARIN-DSPACE repository at LINDAT/CLARIN. *Grey Journal (TGJ)*, *16*.

Xu, X., Arnardóttir, Þ., & Ingason, A. K. (2023). A multilingual database for Icelandic l2 flashcards. *CLARIN Annual Conference Proceedings*, 168–172.

New laws, new opportunities – the effect of the Digital Services Act and the Data Act on access to language data for research purposes

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

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

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

Andrius Puksas
Vytautas Magnus University
Lithuania
andrius.puksas@vdu.lt

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

Benito Trollip
SADiLaR, South Africa
benito.trollip@nwu.ac.za

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

Abstract

This abstract discusses the impact of the Digital Services Act (Regulation 2022 (EU) 2022/2065 of 19 October 2022) and of the Data Act (Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023) on access to language data for research purposes. The referred legal acts significantly contribute to the existing regulatory infrastructure for language research at the EU level.

1 Introduction

In the past couple of years, we have witnessed unprecedented legislative activity from the European Union regarding data access and re-use. In February 2020, the European Commission launched the European Strategy for Data (European Commission, 2020) aimed at creating a single market for data and ensuring Europe's global competitiveness and data sovereignty. The creation of Common European Data Spaces was part of this Strategy. According to the Commission, Common European Data Spaces will ensure that more data becomes available for use in the economy and society, while keeping the companies and individuals who generate the data in control.

In order to achieve the goals of this Strategy, the Commission proposed a series of regulations: Data Governance Act, Digital Services Act, Digital Markets Act, Artificial Intelligence Act and Data Act. All of these have now been enacted into law as EU Regulations. A previous publication (Kamocki et al., 2023) discusses the Data Governance Act and its impact on CLARIN. This submission will discuss the impact of the Digital Services Act (DSA) and of the Data Act (DA) on language research. The authors argue that the referred legislation reconceptualises the framework for accessing language data without interfering with the legal status of language data (see, e.g., Recital 13 of DA).

2 Data Act

The Data Act was adopted on 13 December 2023 and published in the first weeks of 2024. As an EU Regulation with EEA relevance, it shall apply in all EEA Member States, i.e., all EU Member States as well as Iceland, Norway (both being non-EU CLARIN members) and Liechtenstein. It shall also apply directly, i.e. without any national transposition (required for EU Directives), in principle starting from 12 September 2025. The implications of the Data Act for South Africa, a non-EU and non-EEA CLARIN member, still have to be explored. It remains to be seen if South Africa would draft similar legislation to facilitate collaboration with the EU, in following what Bradford (2020) calls the Brussels effect.

From the perspective of CLARIN, and the language community in general, the most immediately relevant parts of the Data Act seem to be Chapters II (on access to and sharing of data from connected devices) and V (enabling public bodies to access private business data in cases of "exceptional need").

According to Chapter II, the user of a connected device ("connected product") should be able to access and share (port) the data generated by the use of the device. A "connected product" is defined as "an item that obtains, generates or collects data concerning its use or environment and that is able to communicate product data via an electronic communications service, physical connection or on-device access, and whose primary function is not the storing, processing or transmission of data on behalf of any party other than the user" (Article 2, (5)). Importantly for language research, this definition includes, as per Recital 23, virtual assistants, such as voice assistants. The data concerned by the above-mentioned access and portability rights are a) product data, i.e. "data generated by the use of a connected product that the manufacturer designed to be retrievable (...)" and b) related service data, i.e. "data representing the digitisation of user actions or of events related to the connected product, recorded intentionally by the user or generated as a by-product of the user's action during the provision of a related service by the provider". Voice input data from voice assistants will likely be included.

According to Article 5 of the Data Act, the user of a connected product can request that the above-mentioned categories of data (both "product data" and "related service data"), together with necessary metadata, be transferred by the data holder directly to a third party, "in a comprehensive, structured, commonly used and machine-readable format and, where relevant and technically feasible, continuously and in real-time" (portability right). Specifically, the "third party" can be a "research organisation" (Recital 33). Personal data related to persons other than the user (e.g., the voice of the user's child or guest captured by a voice assistant), and data containing trade secrets, are partly excluded from this framework. Recital 30 of the Data Act emphasises furthermore that "Any intellectual property rights should be respected in the handling of the data".

The third party that receives the data (e.g., a research organisation) is obliged to use the data "only for the purposes and under the conditions agreed with the user", and "erase the data when they are no longer necessary for the agreed purpose, unless otherwise agreed with the user" (Article 6). The data should not be made available to anyone else, unless the user agrees to it, and the new recipient accepts all the conditions for accessing the data.

In principle, the third party that the user shares the data with shall pay the data holder (the provider of the connected device, not the user) a "reasonable compensation". However, if the third party is a not-for-profit research organisation, this compensation should not exceed the costs incurred by the data holder.

This framework can potentially become an interesting source of voice data; volunteers could directly "donate" their voice data to a research organisation for a specific project.

Another interesting provision of the Data Act refers to the right of public sector bodies to request access to private business data in case of exceptional need (Chapter V). An exceptional need is a situation where a public sector body needs the data to respond to a "public emergency" and cannot obtain the data by alternative means in a timely and effective manner (Article 15). A "public emergency" is defined as "an exceptional situation, limited in time, such as a public health emergency, an emergency resulting from natural disasters, a human-induced major disaster, including a major cybersecurity incident, negatively affecting the population of the Union or the whole or part of a Member State, with a risk of serious and lasting repercussions for living conditions or economic stability, financial stability, or the substantial and immediate degradation of economic assets in the Union or the relevant Member State and which is determined or officially declared in accordance with the relevant procedures under Union or national law".

The data obtained by public sector bodies (the Commission, state or local authorities) can then be shared with research organisations to carry out not-for-profit research that is compatible with the purpose for which the data was obtained. One can imagine a scenario (an exceptional one, admittedly) in which this becomes a source of data for the language community (e.g., tweets related to a natural disaster or a major cybersecurity incident).

3 Digital Services Act

The Digital Services Act (DSA) was adopted on 19 October 2022. It became applicable on 17 February 2024, so now it applies directly in all EEA Member States. Although the main purpose of the DSA is to amend the rules introduced by the eCommerce Directive of 2000 (regarding, most prominently, the liability of providers of intermediary services), it also contains a wide range of other provisions. One of those, Article 40, relates to access to and scrutiny of data held by providers of very large online platforms and very large search engines, and it can be relevant for the language research community.

As per Article 33 of the DSA, a very large online platform is a platform that has a number of average monthly active [users] in the Union equal to or higher than 45 million. These platforms are specifically designated by the Commission[1]. Currently, 17 platforms are designated, including those that are a popular source of data for language research, i.e.: Facebook, Twitter/X, Wikipedia or YouTube. The Commission has also identified two very large search engines: Google and Bing.

Digital Services Coordinators (DSCs) play a key role in the discussed framework. DSCs are competent authorities for all matters related to DSA compliance (Article 49 DSA). Every Member State should have appointed one or more DSCs by 17 February 2024, but some States still have not done it[2]; as a general rule, it seems that existing (tele-) communication authorities are appointed as DSCs.

A researcher who would like to access data from the above-mentioned very large online platforms should first submit an application either directly to the DSC "of establishment", (i.e. the DSC of the country where the provider of the very large online platform has its main establishment in the EU), or to the DSC of the country where his or her research organisation is established (DSC "of the Member State"). In the latter case, the DSC "of the Member State" will first conduct an initial assessment, and then transmit the application to the DSC "of establishment" who will decide whether the applicant can be granted the status of a "vetted researcher" for the specific project described in the application.

A "vetted researcher" should meet all of the following criteria (Article 40(8) DSA):

- be affiliated to a research organisation within the meaning of the DSM Directive, i.e., essentially, a not-for-profit one;
- be "independent from commercial interests";
- disclose funding of the research in the application;
- be capable of fulfilling the specific data security and confidentiality requirements to protect personal data; the application should describe the appropriate technical and organisational measures put in place to this end;
- demonstrate in the application that access to the data requested is necessary for, and proportionate to, the purposes of his or her research, and that the expected results of that research will contribute to the detection, identification and understanding of *systemic risks* in the Union (see below), and to the assessment of the adequacy, efficiency and impact of the risk mitigation measures;
- carry out research for the purposes mentioned in the previous point;
- commit to making the research results publicly available free of charge, within a reasonable period after the completion of the research, subject to the rights and interests of the recipients of the service concerned, in accordance with the GDPR.

Systemic risks are defined in Article 34(1) DSA as including (but, technically, not limited to):

- the dissemination of illegal content through very large online platforms or very large online search engines;
- any actual or foreseeable negative effects (of very large online platforms) for the exercise of fundamental rights (as defined in the EU Charter of Fundamental Rights), *in particular* to human dignity, private and family life, freedom of expression and information, including freedom and pluralism of the media, non-discrimination, child rights and consumer protection;

[1] https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413 (last access: 05.09.2024)

[2] https://digital-strategy.ec.europa.eu/en/policies/dsa-dscs (last access: 05.09.2024)

- any actual or foreseeable negative effects on civic discourse and electoral processes, and public security;
- any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being.

In our opinion, it takes just a little creativity to demonstrate that language research using social media data can contribute to detecting, identifying and understanding of systemic risks. Technically, it seems that a "vetted researcher" should carry out research that also assess the adequacy, efficiency and impact of systemic risk mitigation measures adopted by very large online platforms (as defined in Article 35 DSA), but it is not entirely clear at this stage if this conjunction (the use of *and* in letter (e) above) should not be interpreted as an alternative, which would seem, in our opinion, quite reasonable.

The DSC "of establishment" notifies the European Commission and the newly created European Board for Digital Services[3] about each request it receives, and notifies the names and contact information, and research purpose of every "vetted researcher" to the Board.

After recognising the status of a "vetted researcher", the DSC "of establishment" shall issue a reasoned request for data access to the provider of a very large online platform or a very large online search engine. The provider cannot refuse; it only has 15 days to ask the DSC to amend the request if the provider does not have access to the requested data or if giving access to the data would lead to significant vulnerabilities in the security of the service or the protection of confidential information. This request for amendments should propose specific modifications (alternative means to access the data, additional security measures, etc.). The DSC then has 15 days to decide on the provider's request.

Subsequently (i.e., after no later than 30 days following the DSC's request for access), the provider should provide access to the requested data without undue delay, "through appropriate interfaces specified in the request, including online databases or application programming interfaces" (Article 40(7)).

Access to the data by a "vetted researcher" is then supervised by the DSC "of establishment", who can terminate the access if it determines that the "vetted researcher" no longer meets the condition for this status.

The Commission, in consultation with the Board, can adopt delegated acts laying down the specific conditions under which such sharing of data with researchers can take place in compliance with the GDPR, as well as relevant objective indicators, procedures and, where necessary, independent advisory mechanisms in support of sharing of data.

4 Conclusion

Both the DSA and the Data Act can, in theory, facilitate access to language data for research purposes. Their main merit is that they seem to shift the burden of complying with relevant copyright legislations and the GDPR away from the user, towards businesses and supervisory authorities. On the other hand, both frameworks may appear rather cumbersome (as they involve signing an agreement with the data holder and paying a fee, or going through a procedure with a competent authority), which may paradoxically make them rather inaccessible for individual researchers, especially less experienced ones, and smaller research projects.

The DSA framework has just become applicable, and the Data Act will become applicable in more than a year, so it is too early to judge their real utility for language research. Nevertheless, the language community needs to be aware of these new mechanisms.

The entire legal infrastructure relevant to language research encompasses, in addition to the DSA and the DA also, the Artificial Intelligence Act, the General Data Protection Regulation and numerous EU directives on intellectual property.

References

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World.* Oxford University Press.

[3] https://digital-strategy.ec.europa.eu/en/policies/dsa-board, last access 26.04.2024

Data Act. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act). Available at https://eur-lex.europa.eu/eli/reg/2023/2854/oj (last access: 26.04.2024).

Digital Service Act. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). Available at https://eur-lex.europa.eu/eli/reg/2022/2065/oj (last access: 26.04.2024).

European Commission (2020). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European strategy for data.* Brussels, 19.2.2020. COM(2020) 66 final.

GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (last access: 26.04.2024).

Kamocki, P., Linden , K., Puksas , A. & Kelli , A. (2023). EU Data Governance Act: Outlining a Potential Role for CLARIN. *Selected papers from the CLARIN Annual Conference 2022*. Linköping Electronic Conference Proceedings, no. 198 , CLARIN ERIC , Utrecht , pp. 57-65. https://doi.org/10.3384/ecp198006 .

Constructing the CLARIAH-EUS CLARIN B-Centre: First Steps

First Author
Department (optional)
University of City, Country
email@domain

Second Author
Department (optional)
University Name without city
City, Country
email@domain

Third Author
Department (optional)
University of City, Country
email@domain

Fourth Author
Department (optional)
University Name without city
City, Country
email@domain

Abstract

This extended abstract details the first steps taken to establish the CLARIAH-EUS CLARIN B-centre repository at HiTZ, the Basque Center for Language Technology. The discussion is divided into two sections that describe the various approaches that have been presently debated, selected, and implemented. These include practical and technical decisions regarding infrastructure, operating system, metadata harvesting, and user interface. The ultimate objective of this brief article is to serve as an updated reference point for the creation of future B-centres within both Spain's CLARIAH-ES infrastructure and other CLARIN affiliated centers.

1 Introduction

The Common Language Resources and Technology Infrastructure (CLARIN) is a European Research Infrastructure Consortium (ERIC) that supports research in the humanities and social sciences by facilitating access to multimodal digital language data and tools. Designed as a distributed infrastructure, CLARIN fulfills its mission largely through the services and sources provided by its affiliated centers. These centers are dedicated to distinct areas of expertise and are structured according to one of the CLARIN-designated types.

The original CLARIN B-centres, the type of center discussed in this article, were first certified just over a decade ago. Since then, more than twenty centers of this kind have been recognized throughout Europe. Their primary function is to offer researchers access to digital resources, services, and knowledge in a stable and dependable manner. The caliber of this important role within the wider infrastructure is ensured through strict criteria that are assessed prior to receiving certification as a CLARIN B-centre (Wittenburg et al., 2018). One result of this process is that a set of common technical requirements are put in place at each center that guarantee users continued access to functioning tools and services. Another equally significant outcome is that individual centers are well-equipped to provide resources that cater to the research interests of their particular community.

The need to establish a repository that can deliver digital language technology to those engaged in Basque-language research and Basque Studies is part of the motivation behind the creation of the CLARIAH-EUS B-centre at HiTZ, the Basque Center for Language Technology. Organizationally speaking, CLARIAH-EUS (Alkorta et al., 2024) is a node within CLARIAH-ES (Carreras et al., 2024), Spain's distributed infrastructure for CLARIN and the Digital Research Infrastructure for the Arts and Humanities (DARIAH). In terms of community, however, CLARIAH-EUS is centered on language rather than territory or region, making it transnational in scope. Our hope is that the CLARIAH-EUS B-centre, currently under construction, will help meet the network's desire to foment a collaborative space for those working with Basque or on Basque-related projects.

The brief discussion that follows will outline some of the initial steps taken towards building the CLARIAH-EUS B-centre to this point, as well as highlight actions and decisions that have yet to be made. In similar fashion to descriptions provided about building B-centres in Portugal (Gomes et al., 2018) and the Netherlands (Broeder et al., 2017), ideally this inside look at our work in progress will serve others in their efforts to launch future B-centres within both Spain's CLARIAH-ES infrastructure and other CLARIN affiliated centers.

2 Requirements

As alluded to above, several requirements must be fulfilled to be certified as a CLARIN B-centre. These are compliant with CoreTrustSeal standards (Standards & Board, 2022), which provide a framework for building and maintaining trustworthy digital repositories that ensure the long-term accessibility and usability of digital data and metadata. Through this framework repositories are expected to provide or observe the following:

- **Purpose and Mission**. The B-centre must specify the types of assets that it will manage and the community that it will serve.

- **Reliability**. The repository is required to implement a robust and reliable system to support its data storage, preservation, and access needs.

- **Quality**. The center should possess clear criteria for accepting data and ensuring its relevance, understandability, and adherence to defined standards. This includes assuring the accurate provenance, authenticity, and integrity of digital objects.

- **Security**. It is essential for the repository to put in place a strategy that will allow it to maintain access and preserve data in the face of unforeseen events. Similarly, it must develop a long-term plan that outlines how it will safeguard digital objects and their associated metadata. This includes secure storage solutions that prevent unauthorized access or modification, as well as comprehensive security measures like firewalls, intrusion detection systems, and access controls.

- **Legal and Ethical Propriety**. The B-centre must establish clear procedures for handling permissions and restrictions with respect to deposited data. It will operate within legal and ethical frameworks, demonstrating responsible data handling practices.

- **Governance**. The center must demonstrate is has sufficient funding and staffing to support the repository's long-term viability. In addition, its management team must possess the necessary expertise to handle the complexities of digital data management.

HiTZ is currently working to put these guidelines and requirements in place. To begin with, we have defined our main objectives and mission in keeping with the goals of CLARIAH-EUS. Secondly, we have assigned staff to oversee the repository's construction and future management, with the expectation that we will hire and train new staff members in the near future. Qualified technicians from HiTZ are currently building the repository's infrastructure, housed at the University of the Basque Country's Department of Computer Science in San Sebastián, Spain.

With respect to more technical questions, we are in the process of gathering information about data management, rights, legality, quality, suitability, and reliability to determine which measures and protocols are best suited to ensure CoreTrustSeal standards are met. For the moment, we have identified and are addressing four main requirements during our implementation:

(1) An SSL certificate. A SSL certificate has been obtained and a web server and java application server have been configured to use it.

(2) Federated Identity Management (FIM[1]) for user identification and authentication when accessing protected resources. We have contacted the Spanish national federation SIR (Sistema de Interconexión de Registros) and are configuring SAML2-based FIM to participate in the CLARIN Service Provider Federation (SPF).

(3) Metadata that conforms to the format utilized by the Component Metadata Infrastructure (CMDI[2]), harvestable via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Clarin DSpace has the necessary tools to allow access to CMDI via OAI-PMH and this will be configured to make them accessible.

(4) Associate Persistent Identifiers (PIDs)[3] for metadata and resources. A Handle local server may be used. While this is complex and must be maintained, it allows for greater flexibility. On the other hand, it is also possible to use an external handle service on the net through an API. This is much easier, but not as flexible. We believe utilizing a external service will be more suitable. As such, we will employ ePIC as external handle service to familiarize ourselves with the configuration process, since we intend to utilize DOI (available with Clarin DSpace 7).

3 The B-Centre Server

This section delves into the creation of the B-centre server, including its infrastructure, operating system, installation procedures, and the required Digital Resource Manager in order to connect with other CLARIN repositories.

3.1 Infrastructure

The infrastructure must be configured to withstand accidental and temporary outages or crashes. This allows users and applications to continue operating without interruption and to access data and services. In our current setup, HiTZ has two servers available that we will configure in failover mode: one active and one passive, ready to take over if the first fails. However, to obtain a better High Availability (HA), we plan to integrate a third server for enhanced resilience and to strengthen our ability to withstand disruptions. Leveraging technology that links three servers, such as Ceph, can offer superior reliability compared to using only two servers.

3.2 Operating System

An operating system (OS) that functions with infrastructures based on two servers must be examined and selected. An OS that allows for straightforward management is ideal. There are several options and we have chosen Proxmox. This choice ensures that all team members responsible for handling the group's infrastructure can easily manage the system. Nonetheless, it is advisable that a primary individual is designated to oversee the process.

3.3 Installation

After selecting the OS, we proceeded with the installation process and configured the cluster following these steps: (i) create the cluster and join the servers; (ii) configure Corosync to work with 2 nodes; (iii) connect servers with a straight cable at 10 Gbps; (iv) configure a zfs PoolStorage; (v) install guest machine VM (Ubuntu server LTS); (vi) implement replication; and (vii) enable High Availability to migrate guest VM on failure.

3.4 Digital Resource Manager

Each CLARIN center is required to maintain data and software, necessitating the establishment of a repository. While various Digital Resource Managers are available, CLARIN does not mandate a specific choice. A common option among centers is the utilization of open source platforms like DSpace. We have

[1] FIM: https://https://www.clarin.eu/content/federated-identity

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

[3] PIDs: https://www.clarin.eu/content/persistent-identifiers

choosen to use the CLARIN DSpace implementation[4] to meet CLARIN requirements and to facilitate connection to the Virtual Language Observatory (VLO). As a starting point Clarin DSpace 5[5] will be used, while waiting for the fully developed Clarin DSpace 7. However, Clarin DSpace 5 provides a solid foundation for initial testing and experimentation.

4 Conclusions and Future Work

The creation of the CLARIAH-EUS B-centre at HiTZ, the Basque Center for Language Technology, is driven in part by the need for a central repository of digital language technology resources specifically designed for Basque research and Basque Studies. CLARIAH-EUS prioritizes language over territory, creating an transnational infrastructure that aims to fulfill the network's vision: a collaborative space for anyone working with Basque or Basque-related projects. We hope that the CLARIAH-EUS CLARIN B-centre will be a cornerstone that will unite and activate the community dedicated to Basque studies. This glimpse into CLARIAH-EUS's ongoing construction details the initial steps taken and highlights areas where decisions are still being made. We hope this transparent approach will be a valuable resource for those involved in launching B-centres within Spain's CLARIAH-ES network and within the wider CLARIN infrastructure.

Acknowledgments

References

Alkorta, J., Farwell, A., Fernandez de Landa, J., Altuna, B., Estarrona, A., Iruskieta, M., Arregi, X., Goenaga, X., & Arriola, J. M. (2024). Clariah-eus: A cross-border clariah node for the basque language and culture. *SEPLN-CEDI2024: Seminar of the Spanish Society for Natural Language Processing at the 7th Spanish Conference on Informatics*.

Broeder, D., Bakker, J. T., Van der Laan, M., Kemps-Snijders, M., Windhouwer, M., & Grootveld, M. (2017). Building clarin infrastructure in the netherlands. *CLARIN in the low countries*, 45–59.

Carreras, F. J., Estarrona, A., Farwell, A., Iruskieta, M., Marco, M., Melero, M., Montejo-Ráez, A., Riaño, D., Rigau, G., Romero, D., Ros, S., Sánchez, E., & Sousa, X. (2024). Clariah-es: Strategic network for the integration in the european research infrastructures in social sciences and humanities. *SEPLN-CEDI2024: Seminar of the Spanish Society for Natural Language Processing at the 7th Spanish Conference on Informatics*.

Gomes, L., Branco, F. A. R., Silva, J., & Branco, A. (2018). Setting up the portulan/clarin repository. *CLARIN Annual Conference 2018*, 108.

Standards, C., & Board, C. (2022, September). CoreTrustSeal Trustworthy Digital Repositories Requirements 2023-2025 Extended Guidance. https://doi.org/10.5281/zenodo.7051096

Wittenburg, P., Van Uytvanck, D., Zastrow, T., Strak, P., Broeder, D., Schiel, F., Boehlke, V., Reichel, U., & Offersgaard, L. (2018). *Clarin b centre checklist* (tech. rep.). Technical Report CE-2013-0095, CLARIN ERIC.

[4] Clarin DSpace implementation: https://github.com/ufal/clarin-dspace

[5] Clarin DSpace installation https://github.com/ufal/clarin-dspace/wiki

FAIR Tool Discovery: an automated software metadata harvesting pipeline for CLARIAH

Maarten van Gompel
KNAW Humanities Cluster
Amsterdam, the Netherlands
`proycon@anaproy.nl`

Abstract

We present the Tool Discovery pipeline, a core component of the CLARIAH infrastructure in the Netherlands. This pipeline harvests software metadata from the source, detects existing heterogeneous metadata formats already in use by software developers, and converts them to a single uniform representation based on schema.org and codemeta. The resulting data is then made available for further ingestion into other user-facing catalogue/portal systems.

1 Introduction

For scholars it is important to be able to find and identify tools suitable for their research, we call this process *tool discovery*. We define *tool* here and throughout this paper to broadly refer to any kind of software, regardless of the interface it offers and the audience it targets. The scholar's requirement to find tools is reflected in the letter F for *Findable* in the ubiquitous acronym FAIR [1] that has received a lot of attention in recent years in academic circles. The term is often adopted to promote quality and sustainability in research software (Jiménez et al., 2018). In order to find tools, researchers must have access to catalogues that relay *accurate* software metadata.

There is no shortage in existing initiatives in building such catalogues[2]. , but the system we describe in this paper is not an attempt to build another catalogue. We developed a generic pipeline that harvests software metadata from the source, leveraging various existing metadata formats, and converting those to a uniform linked open data representation. This data can then be used to feed catalogues.

2 The need for high-quality metadata

Unlike most digital data, software is uniquely characterised as a constantly moving target rather than a static deliverable entity. Releases address bugs, security vulnerabilities, or add new features. Moreover, software lives not in isolation, but in connection to other software; its dependencies. Updates are needed to adapt to changes in its runtime environment.

For software metadata to be informative in this dynamic setting, it needs to reflect this moving target and explicitly link to a particular version of the software. This also facilitates provenance keeping and scientific reproducibility. Metadata should also convey information about the stage of development the software is in and the level of support an end-user may expect. The user would be wise to exercise caution in adopting software that is unmaintained and unsupported. In practice we often find this information lacking and come across catalogues that were manually compiled once but rarely updated since.

The need for accurate up-to-date metadata goes hand-in-hand with the need for *complete* metadata. If vital details are missing, the end-user may not be able to make an informed judgment.

[1] Findable, Accessible, Interoperable and Reusable

[2] CLARIN itself has one in the form of the Virtual Language Observatory (VLO): https://vlo.clarin.eu (Uytvanck et al., 2010)

3 Bottom-up harvesting from the source

What we propose is a *fully automated* pipeline where software metadata is kept at the source, i.e. alongside the software source code, and *periodically* harvested from there. This is in contrast to approaches where metadata primarily resides in an intermediate system that is manually constructed or curated, which is a common approach for many software catalogues[3]. Our approach has a number of important advantages:

Source code is often already accompanied by software metadata in existing schemas[4] because many programming language ecosystems already either require or recommend this. Our aim is to avoid any duplication of metadata and *reuse* these existing sources to the maximum extent possible.

Second, source code should be under version control (e.g. git) and published in a forge (e.g. GitHub, Gitlab, Sourcehut). This solves versioning issues and ensures metadata can exactly describe the version alongside which it is stored. It also enables the harvester to properly identify the latest stable release[5]. Software forges themselves may also provide an API that may serve as an extra source to find software metadata (e.g. descriptions, keywords, links to issue trackers and/or continuous integration services). Third, the developers of the tool have full control and authorship over their metadata. There are no middlemen. Last, the forges were designed precisely for collaboration on open source software development, so mechanisms for any third party to amend or correct the metadata are already in place (e.g. via a pull/merge request or patch via e-mail). So while developers retain full authorship, this does not mean outside contribution and curation is not possible.

We do not harvest any metadata from intermediaries[6] as that would defeat our philosophy. We do have one extra source for harvesting: In case the tool in question is Software as a Service, i.e. a web-application, web-service, or website, we harvest not only its software source code, but also its web endpoint and attempt to automatically extract metadata from there[7]. In the resulting metadata, there will be an explicit link between the source code and any *target products* or *instances* of that source code. The sources for harvesting source repositories and web endpoints (URLs) are the only input that needs to be manually provided to our system, we call this the *source registry* and we keep this in a simple git repository containing very minimalistic configuration files (one yaml file per tool). This is also the only point in our pipeline where there is the possibility for a human curator to decide whether or not to include a tool.

4 A unified vocabulary for software metadata

The challenge we are facing is primarily one of mapping from multiple heterogeneous sources of software metadata to a unified vocabulary. Fortunately, this is an area that has been explored previously in the CodeMeta project[8]. They developed a vocabulary for describing software source code, extending the schema.org vocabulary and contributing their efforts back to them. Moreover, the CodeMeta project defines mappings, called crosswalks, between their vocabulary and many existing metadata schemes.

Schema.org and codemeta are both linked open data (LOD) vocabularies[9], and codemeta is canonically serialised to a JSON-LD[10] file which makes it easily parsable for both machine and human alike. This `codemeta.json` file can be kept under version control alongside a tool's source code. The developer has a choice to either run our harvester and converter themselves and commit the resulting codemeta file, or to not add anything and let the harvester dynamically reconstruct the metadata every harvest cycle.

We link to various other LOD vocabularies, such as *repostatus.org* (development status), *SPDX* (open

[3] for example, https://research-software-directory.org/ offers such a platform. Metadata can often be exported via OAI-PMH.

[4] for example `pyproject.toml`, `setup.py`, `pom.xml`, `package.json`, `CITATION.cff` and others. Even files such as `README` and `LICENSE` may be a source for certain metadata.

[5] provided some kind of industry-standard versioning system like semantic versioning is adhered to

[6] from other catalogues, for instance via the aforementioned OAI-PMH endpoints

[7] Formally, the software source code has no knowledge when, where, and by whom it may be deployed as a service. This linking is therefore established at a higher level.

[8] https://codemeta.github.io

[9] i.e. building upon RDF and being retrievable over HTTP

[10] https://www.w3.org/TR/json-ld/

source software licenses), *TaDiRaH* (research activities) (Borek et al., 2016) and *NWO Research Domains*[11]. Moreover, we formulated some of our own extensions on top of codemeta and schema.org, such as Software Types and Software Input/Output Data[12], as well as Research Technology Readiness Levels. Most of these are formulated as SKOS[13] vocabularies.

5 Architecture

The full architecture of our pipeline is illustrated schematically in Figure 5. Although we demonstrate this in the context of the CLARIAH project, the underlying technology is generic and can also be used for other projects.

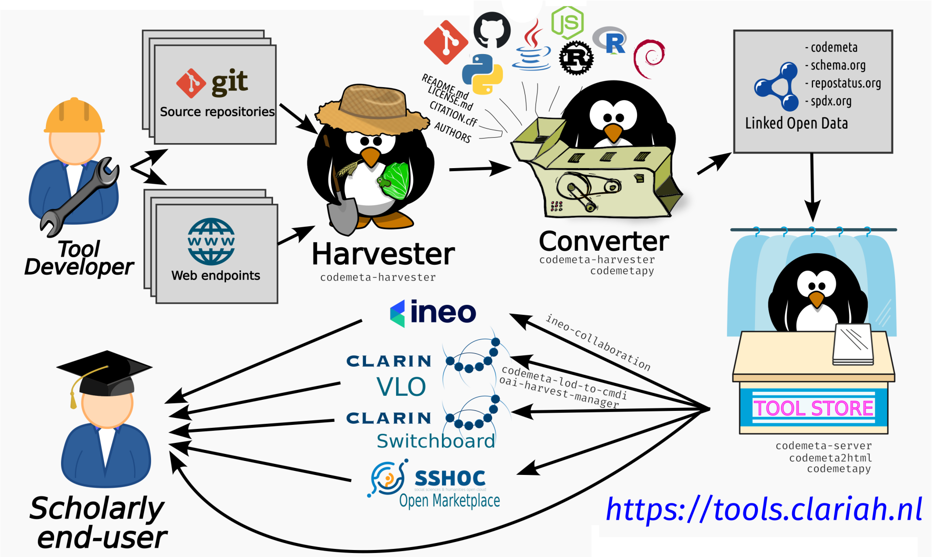


Figure 1: The architecture of the CLARIAH Tool Discovery pipeline

Using the input from the source registry, our *harvester*[14] fetches all the git repositories and queries any service endpoints. It does so at regular intervals (e.g. once a day). This ensures the metadata is always up to date. When the sources are retrieved, it looks for different kinds of metadata it can identify there and calls the converter[15] to turn and combine these into a single codemeta representation. This produces one codemeta JSON-LD file per input tool. All of these together are loaded in our *tool store*. This is implemented as a triple store and serves both as a backend to be queried programmatically using SPARQL, as well as a simple web frontend to be visited by human end-users as a catalogue [16]. The frontend for CLARIAH is accessible at https://tools.clariah.nl, with at the time of writing 114 registered source repositories and 34 web endpoints.

Our web front-end is not the final destination; our aim is to propagate the metadata we have collected to other existing portal/catalogue systems, such as the CLARIN VLO, the CLARIN Switchboard, the SSH Open Marketplace, and CLARIAH's Ineo[17]. The latter has already been implemented, the VLO export will be done via a conversion from codemeta to CMDI, and the Marketplace conversion has started in collaboration with DARIAH.

[11] https://repostatus.org, https://spdx.dev, https://vocabs.dariah.eu/tadirah/, https://www.nwo.nl/en/nwo-research-fields

[12] https://github.com/SoftwareUnderstanding/software_types, https://github.com/SoftwareUnderstanding/software-iodata

[13] https://www.w3.org/TR/skos-reference/

[14] codemeta-harvester: https://github.com/proycon/codemeta-harvester

[15] powered by codemetapy: https://github.com/proycon/codemetapy

[16] powered by codemeta-server (https://github.com/proycon/codemeta-server) and codemeta2html (https://github.com/proycon/codemeta2html).

[17] https://vlo.clarin.eu/, https://switchboard.clarin.eu/, https://marketplace.sshopencloud.eu/, https://ineo.tools

6 Validation & Curation

Having an automated metadata harvesting pipeline may raise some concerns regarding quality assurance. Data is automatically converted from heterogeneous sources and immediately propagated to our tool store, this is not without error. In absence of human curation, which is explicitly out of our intended scope, we tackle this issue through an automatic validation mechanism.

The harvested codemeta metadata is held against a validation schema (SHACL) that tests whether certain fields are present (completeness), and whether the values are sensible (accuracy, it is capable of detecting various discrepancies). The validation process outputs a human-readable validation report which references a set of carefully formulated *software metadata requirements* [18]. Developers can clearly identify what specific requirements they have not met. The over-all level of compliance is expressed on a simple scale of 0 to 5, and visualised as a coloured star rating in our interface. This evaluation score itself is part of the delivered metadata and something which both end users as well as other systems can filter on. It may even serve as a kind of 'gamification' element to spur on developers to provide higher quality metadata. We find that human compliance remains the biggest hurdle and it is hard to get developers to provide metadata beyond what we can extract automatically from their existing sources. For CLARIAH we measure: 5 stars (2%), 4 (23%), 3 (45%), 2 (7%), 1 (19%), 0 stars (4%).

For propagation to systems further downstream, we set a threshold rating of 3 or higher. Downstream systems may of course posit whatever criteria they want for inclusion, and may add human validation and curation. As metadata is stored at the source, however, we strongly recommend any curation efforts to be directly contributed upstream to the source, through the established mechanisms in place by whatever forge (e.g. GitHub) they are using to store their source code.

7 Conclusion

We have shown a way to store metadata at the source and reuse existing metadata sources, recombining and converting these into a single unified LOD representation using largely established vocabularies. We developed tooling for codemeta that is generically reusable and available as free open source software[19]. We hope that our pipeline results in metadata that is accurate and complete enough for scholars to assess their usability for their research. We think this is a viable solution against metadata or entire catalogues going stale, in worst case unbeknownst to the researcher who might still rely on them.

Acknowledgments

The FAIR Tool Discovery track has been developed as part of the CLARIAH-PLUS project (NWO grant 184.034.023), as part of the Shared Development Roadmap.

References

Borek, L., Dombrowski, Q., Perkins, J., & Schöch, C. (2016). Tadirah: A case study in pragmatic classification. *Digit. Humanit. Q.*, *10*(1). http://dblp.uni-trier.de/db/journals/dhq/dhq10.html#BorekDPS16

Jiménez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., Capella-Gutiérrez, S., Hong, N. P. C., Cook, M., Corpas, M., Flannery, M., García, L. J., Gelpi, J. L., Gladman, S. L., Goble, C. A., Ferreiro, M. G., González-Beltrán, A. N., Griffin, P., Grüning, B. A., … Crouch, S. (2018). Four simple recommendations to encourage best practices in research software. https://api.semanticscholar.org/CorpusID:214915242

Uytvanck, D. V., Zinn, C., Broeder, D., Wittenburg, P., & Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Lrec*. European Language Resources Association. http://dblp.uni-trier.de/db/conf/lrec/lrec2010.html#UytvanckZBWG10

[18] CLARIAH Software Metadata Requirements: https://github.com/CLARIAH/clariah-plus/blob/main/requirements/software-metadata-requirements.md

[19] GNU General Public Licence v3

A core metadata schema for interpreting corpora: Implementation on the Unified Interpreting Corpus (UNIC) platform

Nannan Liu
University of Bologna, Italy
nannan.liu@unibo.it

Mariachiara Russo
University of Bologna, Italy
mariachiara.russo@unibo.it

Abstract

This study presents a metadata schema for interpreting corpora and the Unified Interpreting Corpus (UNIC; https://unic.dipintra.it/) platform. The two infrastructures were created based on the FAIR (findability, accessibility, interoperability, and reusability; Wilkinson et al., 2016) and CARE (collective benefit, authority to control, responsibility and ethics; Carroll et al., 2020) principles of scientific data management, a review of 125 interpreting corpora, and open-source tools. We describe our plans for gathering user feedback in early October.

1 Introduction

Conceptualised as ‘speech-to-speech translation’ and ‘signing’ in lay discourse, interpreting refers to immediate rendition in the target language based on a one-time presentation in the source language (Pöchhacker, 2022). Compared with translation, interpreting corpora interweave multilingualism with multimodality and split-second processing with contextualised interactions. However, both interpreting and translation corpora are missing from CLARIN’s list of resource families[1]. Surveys with translation students indicate that CLARIN’s existing metadata information and functionalities were largely unhelpful to this group of users (Lušicky & Wissik, 2017).

We address three questions: Empirically, how are interpreting corpora described with metadata, namely structured information about a corpus? Conceptually, what should be a metadata schema for interpreting corpora that adheres to FAIR and CARE principles (Carroll et al., 2020; Wilkinson et al., 2016)? Technically, how do we implement the schema on an open-source platform?

2 Methods

We define an ‘interpreting corpus’ as a sample of transcripts, sign annotations, and/or audiovisual recordings selected to represent a target interpreting domain or interpreter population. We focused on parallel corpora, where metadata are “much poorer” in CLARIN (Fišer et al., 2018, p. 4).

2.1 Corpus collection

To identify interpreting corpora, we searched three bibliographies (e.g. Translation Studies Bibliography, Conference Interpreting Research Information Network [CIRIN] Bibliography) and seven repositories (e.g. Virtual Language Repository, Google Dataset Search) and consulted the references of collected documents. We last queried the databases on 30 July 2024.

Based on the conceptual analysis of Pöchhacker (2024) and our explorations, the query terms for non-interpreting-specific databases were ‘interpreting corpus’, ‘interpretation corpus’, ‘signed language translation corpus’, ‘sign language translation corpus’, ‘speech translation corpus’, ‘spoken language translation corpus’, and ‘speech-to-speech translation corpus’. The keywords were ‘corpus’ and ‘corpora’ for the interpreting-specific CIRIN Bibliography. The databases’ bibliographic information of non-English entries was translated into English, so just English query terms were used.

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

This process resulted in a list of 1,898 publications, and we analysed those in Chinese, English, Italian, and Spanish to see if an interpreting corpus was presented or analysed. In five cases of indeterminacy, we discussed the texts to reach a consensus. To understand the reusability of the corpora collected, we coded their attributes, availability, licence, and provenance (Wilkinson et al., 2016). Following Bird and Simons (2003), we documented 'availability' along a scale from 'open' (the full dissemination of data free of charge and without restrictions in use), 'restricted' to specific uses (e.g. research and teaching only) or payment conditions, to 'closed' where data are not publicly available.

2.2 FAIRness assessments

We examined whether the collected corpora have landing pages and publication-internal metadata. Regarding the webpages, we applied the FAIR Evaluator (Wilkinson et al., 2019) to assess six records in six different repositories, which describe the most widely published corpora. We supplied the metadata identifiers to the Evaluator, which executed 14 tests assessing the record's compliance with 11 FAIR principles (Wilkinson et al., 2019).

To identify "accurate and relevant attributes" for reuse (Wilkinson et al., 2016, p. 4), we performed a manual 'overlap and gap' analysis of the publication-internal and unique web metadata in different repositories. Overlaps refer to exact and partially matched, synonymous, and related terms used in more than 50 percent of the metadata collected. Gaps are information that can be gleaned from more than 50 percent of the corpora but not found in the metadata, i.e. what is in the data but not metadata. We compared the results of the overlap analysis and jointly determined the names of the overlaps and gaps, prioritising concepts available in CLARIN registries and shared by different types of corpora.

2.3 Designing the schema and UNIC

Because of their relevance to most corpora, we made the components and elements gathered from overlap and gap analyses compulsory. As regards the optional fields, we examined the (meta)data of four corpora in parliamentary (Russo et al., 2012), media (Liu, 2023), healthcare (Bührig et al., 2012), and signed language settings (Vandeghinste et al., 2022), including terms shared by at least two collections. We devised elements for other settings based on our collection.

We distributed previous versions of the schema at two translation and interpreting conferences and to a corpus-linguistic mailing list (corpora@list.elra.info) to bring forward aspects needing clarification. This abstract presents version 1.0 of the schema implemented on the UNIC platform. Hosted on a web server on a Linux operating system in the authors' department, UNIC uses PostgreSQL to store the (meta)data, the query language SQL to communicate between the database and server, and Astro, React, and Next.JS to build the web interface. All the foregoing tools are open-source.

3 Findings

3.1 Interpreting corpora

We identified 125 corpora in 38 spoken and eight signed languages, which were described and analysed in 382 publications. Regarding availability, 103 corpora are closed, access to 17 is restricted, and just five are open. Whereas three open-access corpora are hosted on the creators' institutional pages, the FerFuLice corpus of interpreting by bilingual twins (Liceras et al., 2008) and the Polish Interpreting Corpus (Chmiel et al., 2021) are distributed by TalkBank and CLARIN-Poland, respectively.

Despite being data silos, the interpreting corpora surveyed have the potential for coalescing around a core metadata set. For instance, the majority described the number of tokens and glosses (65/125), interpreters (81), and duration of audiovisual recordings (81). Statistics available show that they contain the performance of an estimated number of 1,266 interpreters, amounting to approximately 38.85 million tokens and glosses and 5,850 hours of recordings.

The uniqueness of annotation and alignment applied to interpreting vis-à-vis translation corpora necessitates domain-relevant standards. Regarding annotation, although 24 corpora were automatically annotated (e.g. part-of-speech tags), 52 involved significant manual labour, e.g. segmenting speeches (10) and categorising poses (5). Fifty-nine corpora were aligned at textual and temporal levels: 41 teams of creators aligned source and target transcripts by such a unit as a turn or a sentence, 29 synchronised transcripts with audiovisual recordings, and two time-aligned source and target audio to study the lag between the original speech and the interpreter's output.

3.2 Metadata FAIRness

Among the 125 corpora collected, we found 23 landing pages and 29 publication-internal descriptions. Results of the automatic assessment of six web records indicate relative accessibility and interoperability, some findability, but a low degree of reusability. Corpora registered in CLARIN-related repositories were FAIRer than those in the European Language Resources Coordination-SHARE and European Language Resources Association repositories.

Results of the manual analyses revealed 21 mandatory elements (e.g. 'description', 'interpreter status') and four compulsory components, namely 'corpus', 'event' (the communicative situations mediated by interpreters), 'transcriber', and 'annotation'.

3.3 The schema and UNIC

The metadata schema for interpreting corpora v1.0 is available at http://tinyurl.com/intpmetadata. Comprising ten components and 95 elements, it reuses 92 CLARIN-defined concepts and introduces the mediation-specific 'source text', 'interpreter', and 'alignment' components. The v1.0 schema is implemented on the UNIC platform (Figure 1), which offers three ways of sharing (meta)data. First, corresponding to the 'corpus' component, the 'register your corpus' page requests minimal administrative and design information to assist with verification and potential collaboration.

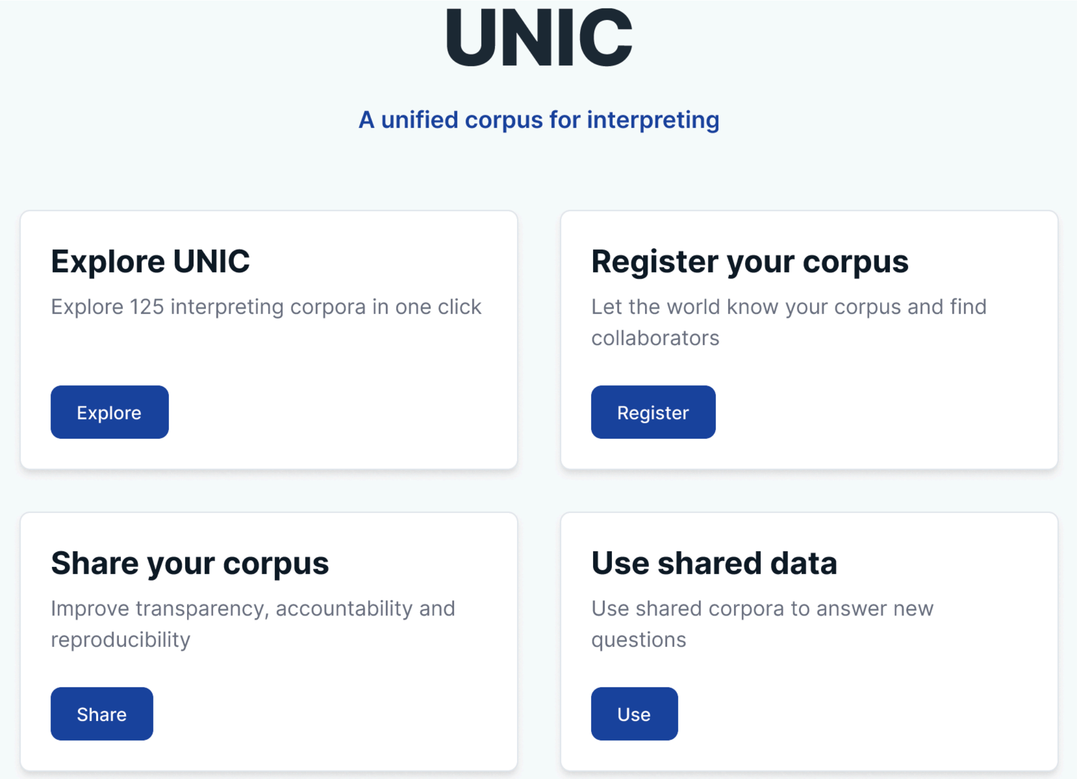

UNIC
A unified corpus for interpreting
Explore UNIC
Explore 125 interpreting corpora in one click
Explore
Register your corpus
Let the world know your corpus and find collaborators
Register
Share your corpus
Improve transparency, accountability and reproducibility
Share
Use shared data
Use shared corpora to answer new questions
Use

Figure 1: The UNIC homepage

Second, at the bottom of the 'register your corpus' page, creators can upload the specific metadata based on the rest of the components using a spreadsheet template that guides them on the schema and automatically validates their input. Third, creators are encouraged to share their corpus data after submitting the metadata. (Meta)data collected from registration, uploading and sharing are available on the searchable ‘Explore UNIC’ page, where mandatory elements other than ‘description’, ‘actor id’, and ‘cite as’ serve as filters. Opt-out mechanisms were devised to protect data subjects’ rights and interests (Carroll et al., 2020).

4 Future work

In early October, we will solicit feedback on the schema and UNIC from 54 interpreting students and about 20 scholars. Following Lušicky and Wissik (2017), we structure the questions around the perceived usefulness of filters, desirable filters, and the degree of satisfaction with UNIC functionalities.

Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101108651.

References

Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, *79*(3), 557–582.

Bührig, K., Kliche, O., Meyer, B., & Pawlack, B. (2012). The corpus ”Interpreting in Hospitals”: Possible applications for research and communication training. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 305–315). Amsterdam: John Benjamins.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., … Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, *19*(1), 1–12. doi:10.5334/dsj-2020-043

Chmiel, A., Janikowski, P., Kajzer-Wietrzny, M., Koržinek, D., & Jakubowski, D. (2021). EU Parliament Speech Corpus. CLARIN-PL digital repository. Retrieved from http://hdl.handle.net/11321/821

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN’s key resource families. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Liceras, J. M., Fernández Fuertes, R., Perales, S., Pérez-Tattam, R., & Spradlin, K. T. (2008). Gender and gender agreement in bilingual native and non-native grammars: A view from child and adult functional–lexical mixings. *Lingua*, *118*(6), 827–851. doi:10.1016/j.lingua.2007.05.006

Liu, N. (2023). Speaking in the first-person singular or plural: A multifactorial, speech corpus-based analysis of institutional interpreters. *Interpreting*, *25*(2), 239–273. doi:10.1075/intp.00088.liu

Lušicky, V., & Wissik, T. (2017). Discovering resources in the VLO: A pilot study with students of translation studies. In *Selected papers from the CLARIN Annual Conference 2016* (pp. 63–75). Linköping: Linköping University Electronic Press.

Pöchhacker, F. (2022). *Introducing interpreting studies* (3rd). London and New York: Routledge.

Pöchhacker, F. (2024). Is machine interpreting interpreting? *Translation Spaces*. doi:10.1075/ts.23028.poc

Russo, M., Bendazzoli, C., Sandrelli, A., & Spinolo, N. (2012). The European Parliament Interpreting Corpus (EPIC): Implementation and developments. In F. Straniero Sergio & C. Falbo (Eds.), *Breaking ground in corpus-based interpreting studies* (pp. 53–90). Bern: Peter Lang.

Vandeghinste, V., Van Dyck, B., De Coster, M., Goddefroy, M., & Dambre, J. (2022). BeCoS corpus: Belgian Covid-19 Sign Language corpus. A corpus for training sign language recognition and translation. *Computational Linguistics in the Netherlands Journal*, *12*, 7–17.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. doi:10.1038/sdata.2016.18

Wilkinson, M. D., Dumontier, M., Sansone, S.-A., Bonino da Silva Santos, L. O., Prieto, M., Batista, D., … Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data*, *6*(1), 174. doi:10.1038/s41597-019-0184-5

Vocabularies in CLARIN : Problems and suggested solutions

Daan Broeder
CLARIN ERIC
Utrecht, the Netherlands
d.g.broeder@uu.nl

Jan Odijk
Utrecht University
Utrecht, the Netherlands
j.odijk@uu.nl

Abstract

We discuss some problems with vocabularies in CLARIN and suggest solutions for them. We illustrate this with concrete examples related to a CMDI profile for services.

1 Introduction

In the past 15 years CLARIN has created a robust and stable metadata infrastructure which mostly serves its current needs. This does not mean that there are no problems. In this paper we want to discuss some:

1. Isolationism: there is a tendency of CMDI isolationism, i.e. not sharing harvested CMDI metadata with larger infrastructures, not allowing references to external concept definitions or vocabularies.
2. Non-functioning of the governance structure around CCR
3. No relations between concepts or other items, in particular no hierarchies . This has led to long flat lists of items in the VLO, CCR, ISOCAT and the Component Registry
4. No support for interfaces that support relations between concepts or other items
5. Component Registry: issues with authorisation, group management, and versioning

In this paper we will describe these problems, especially in the context of a long-term attempt to come to a sufficiently specific CMDI profile for describing services that were created and made available in the Netherlands CLARIAH projects as well as a range of WebLicht Web Services (Hinrichs et al., 2010).[1]

2 CSD Profile for Linguistic Services

Odijk (2019) describes a CMDI (Broeder et al., 2012) profile for the description of software called *ClarinSoftwareDescription*, CSD.[2] This profile was created on the one hand to be able to provide better and more detailed descriptions of software in the CLARIN research infrastructure and on the other hand to make it possible to discover software more easily than was possible at the time using the CLARIN Virtual Language Observatory (VLO).

The profile was used in CMDI descriptions for more than 80 software resources, mainly from the Netherlands, and in more than 265 descriptions for WebLicht web services. As a proof of concept, a search interface (CLAPOP) was created by Daan Broeder to show that discovering and finding software resources in a faceted search application using the properties of the CSD profile was much easier than in the VLO. This search interface, both for the software from the Netherlands and for the combined Netherlands and Weblicht software is unfortunately no longer available due to maintenance and compatibility issues. The hope was that the VLO would be extended with a separate search interface for discovering software resources based on the CSD profile and accompanying metadata records, but that hope turned out to be futile so far.

The faceted search interface uses a number of facets, of which we mention two that are relevant here:

Linguistic Subject given the CLARIN focus on language, linguistics plays a prominent role in it, and this facet enables the user to select software suited for specific subdisciplines within linguistics

Tool Task an indication of the function of the software

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

[2] clarin.eu:cr1:p_1342181139640

3 Linking to the CLARIN Concept Registry (CCR)

During the development of the CSD profile various vocabularies were created, and a proposal was made to link the items in these vocabularies to entries in the CLARIN Concept Registry (CCR)[3] or to create new entries in the CCR where needed. As described in (Odijk, 2019), the incorporation of new entries never happened because the committee to assess the proposal was not functioning at all (Problem 2), in part due to lack of funding, itself in part due to lack of appreciation of the importance of this aspect of a properly functioning research infrastructure.

4 SSH Vocabulary Commons

The SSH Vocabulary Commons[4] is a result of the SSHOC project and collaboration between the partners of the SSH Cluster. The SSH Vocabulary Commons provides a place to host and publish Vocabularies relevant for the SSH that lack support from other organisations.

Since CLARIN does not offer a flexible SKOS vocabulary hosting service, we recently updated some of the vocabularies and their definitions and integrated them in the SSH Vocabulary Commons. We did this in particular for the facets *tool tasks*[5] (121 items) and *linguistic subject*[6] (46 items).

The SSH Vocabulary Commons uses the widely used SKOSMOS software to publish vocabularies specified as SKOS Concept Schemes. The full power of SKOS enables one to specify relations between concepts, in particular one or more hierarchies. This is a long-desired feature, in fact, in our view, a feature without which infrastructures services cannot function properly. This was already pointed out by (Odijk, 2009, pp. 12–13) when discussing ISOCAT, reiterated in (Odijk, 2014, pp. 8, 13) and again reiterated in the context of the current properties in (Odijk, 2019, p. 123). This concerns Problems 3 and 4.

CLARIN followed an active policy of not having a hierarchy in ISOCAT and CCR because this would be controversial for different groups of researchers. Though it is surely the case that any hierarchy will raise controversies, in our view one should have the policy to enable multiple hierarchies (and other relations between concepts) and *be explicit* to have no ontological claims with these hierarchies but only pragmatic concerns to facilitate searching. The absence of relations between items, e.g. in a small hierarchy is in our view an important reason for the, mildly put, less than optimal functioning of services such as ISOCAT, CCR, CLARIN Component Registry, various CMDI editors, and VLO.

The reason is that repeated searching in a long flat list for a possible value is simply not done by users and can also not be reasonably asked from them. The consequence is that the relevant work is not done at all, or that the user creates his/her own new entry that very likely is a duplicate of an already existing item, leading to a proliferation of items for the same concept. Not surprisingly, the CCR contains 3 entries for "noun", 1 for "Noun", and additionally entries for "Noun(plural)" and "Noun(singular)". We happen to see these terms together because they are alphabetically close together. But similar or related terms that start with different letters will not be easily spotted. We tried the term "substantive" and it indeed occurs and refers to one of the concepts labelled "noun". Similarly, there are 3 entries for 'neuter', one for 'neuter gender' and one for 'NeuterGender'.[7]

There were some other properties to subdivide concepts in the CCR, but these involved whole linguistic subdisciplines (e.g. Morphosyntax, Metadata, Terminology, etc.). The most recent web interface to the CCR[8] shows this under the tag 'Groups', but no list of concepts belonging to such groups is (currently?) shown, so this is not very useful so far. This interface also shows a tag 'Hierarchy', but no hierarchy can be shown there because there is none.

The hierarchical relations we specified for the items of these vocabularies can be viewed in the SSH Vocabulary Commons Interface (https://vocabs.sshopencloud.eu/browse/csd-tool-tasks/en/) under the

[3] https://www.clarin.eu/content/clarin-concept-registry.

[4] https://vocabs.sshopencloud.eu/.

[5] https://vocabs.sshopencloud.eu/browse/csd-tool-tasks/en/

[6] https://vocabs.sshopencloud.eu/browse/csd-linguistic-subjects/en/

[7] We admit that another reason for this may be differences between views of researchers on how linguistic concepts should be characterised and subdivided, but we are convinced that this is not the only factor.

[8] Apparently embedded in CLAVAS (Brugman, 2017).

tag ‘Hierarchy’. The major goal of the hierarchy is (1) to group semantically close items together so that they can be found together especially if the user does not know the exact term that is used for a particular concept, and (2) to show at any level in the hierarchy only a small (ideally less than 15, and preferably much smaller) number of items. If the number of items to search for is large, users will not bother searching in it. Of course, though the hierarchy shortens the number of items at each level in the hierarchy, it does create a hierarchy depth, which creates extra effort. But repeated searching for items in small lists through a hierarchy is still much easier and takes less effort than repeated searching in a long list of more than 120 items, especially since the user gets very quickly acquainted with the top level hierarchy. This is well known from Time Complexity Analysis[9] and in the current taxonomy for tools and services tasks the number of items to be inspected to get at a concept in the worst case equals to 22 + 3 + 12 + 6 = 43, about a third of the worst case in a search in an flat list of 120 items.

Note that multiple hierarchies are allowed and possible, and the only criterion for their usefulness is whether they facilitate searching. In fact, in the current vocabulary for tool tasks there are multiple hierarchies: the linguistic *tooltasks* can be reached both via the path *analysis/linguistic analysis* and via the path *annotation/linguistic annotation*. Other examples will be given in the full paper.

5 Relation with TaDiRAH

TADIRAH is a **Ta**xonomy of **Di**gital **R**esearch **A**ctivities in the **H**umanities. Since many research activities are currently by necessity conducted using tools, it is to be expected that there is some overlap between TaDiRAH and the CLARIN Tool and Service Tasks (CSD-TT) taxonomies.

At the DH2023 pre-conference on TaDiRAH it was argued that from a CLARIN perspective, TaDiRAH was mostly too general to be practical for describing typical CLARIN LT services (Borek et al., 2023). This proves a challenge for instance in making the SSH Open Marketplace[10], which uses TaDIRAH to classify services and (workflow) solutions from the broad SSH, also useful for the CLARIN community, and for using TaDiRAH common service classification vocabulary in any common SSH application such as for example the Dutch CLARIAH project catalogue INEO,[11] which uses TaDiRAH in its search interface and provides the definitions of the TaDiRAH concepts.[12]

Several solutions can be envisaged, for example by linking and merging TaDIRAH and CSD-TT into a fully combined taxonomy in a way that respects the current separate editorial responsibilities of the TaDIRAH and CLARIN communities. The two vocabularies should ideally remain available as independent vocabularies. In general the collaboration between the SSH Cluster partners and beyond with other disciplines should motivate to look for common categorization systems provided these can be made specific enough.

Some of the concepts that we use in CSD-TT have an equivalent in TADIRAH. An example is the CSD-TT concept with label *analysis*,[13] which is equivalent to the TADIRAH concept with label *Analyzing*.[14] This equivalence has been made explicit by means of the *skos:exactMatch* property. We will provide more examples in the full paper.

In the full paper we will describe more avenues that are currently being investigated to integrate vocabularies such as CSD-TT and TaDiRAH, which is an ongoing discussion topic in the SSHOC network especially with the colleagues from the DARIAH Vocabulary service.

6 Integrating the taxonomies in the CSD profile

With the relevant concepts and associated labels in the SSH Vocabulary Commons *CSD-TT vocabulary*, it is natural to improve the existing CSD profile by adding links to the relevant concept to the labels used in this profile (which are mostly identical to the preferred labels in the SSH Vocabulary Commons.

[9] https://en.wikipedia.org/wiki/Time_complexity.
[10] https://marketplace.sshopencloud.eu/.
[11] https://www.ineo.tools/
[12] https://www.ineo.tools/definitions/research-activities
[13] https://vocabs.sshopencloud.eu/browse/csd-tool-tasks/en/page/0004
[14] https://vocabs.dariah.eu/tadirah/en/page/analyzing.

However, a problem is that CLARIN does not allow vocabularies other than those provided by CLAVAS, which only contains the ISO-639 vocabulary for languages and the CCR. External vocabularies such as the one we created are not allowed (Problem 1). Furthermore, the CLARIN Component Registry only allows one to add flat lists of items to a metadata element: it is not possible to specify that a particular *remote or local hierarchical structured* vocabulary must be used.

An additional problem with the component registry, already in earlier phases when CSD was being developed and recently again, is that the arrangement of the authorisation of groups of researchers that work together on a metadata profile has not been taken care of properly. Also versioning of profiles and components was not properly arranged. We got excellent support from the developers, who took ad-hoc measures to take care of our problems, but such ad-hoc measures are of course not desirable and should not be necessary (Problem 5)

7 Recommendations

On the basis of our experiences, we conclude with some recommendations that we will elaborate on in the full paper:

- Allow flexible group management and ownership of all types of profiles and components, including easy change of group membership
- Allow for concept definitions to be provided outside the CCR and CLAVAS
- CMDI should support external/remote vocabularies hosted elsewhere.
- Make sure that tools using CMDI such as metadata editors, CCR, Component Registry and VLO can use hierarchies to facilitate finding and selecting values so that re-use instead of proliferation is stimulated
- Have at least one CLARIN supported CMDI editor that supports hierarchical selection of items
- Revise governance around CCR

References

Borek, L., Hastik, C., Dombrowski, Q., Broeder, D., Rockenberger, A., Nagasaki, K., Mochizuki, R., Katakura, S., Cupar, D., & Ohmukai, I. (2023). Multilingual taxonomy initiative - TaDiRAH as community of practice [Digital Humanities 2023. Collaboration as Opportunity (DH2023)]. http://doi.org/doi:10.5281/zenodo.8107850

Broeder, D., Windhouwer, M., Uytvanck, D. V., Goosen, T., & Trippel, T. (2012, May). CMDI: A component metadata infrastructure. In *Proceedings of the workshop describing language resources with metadata: Towards flexibility and interoperability in the documentation of language resources* (pp. 1–4). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf

Brugman, H. (2017). CLAVAS: A clarin vocabulary and alignment service. In J. Odijk & A. van Hessen (Eds.), *Clarin in the low countries* (pp. 61–69). Ubiquity Press. https://doi.org/10.5334/bbi.5

Hinrichs, E. W., Hinrichs, M., & Zastrow, T. (2010). WebLicht: Web-Based LRT Services for German. *Proceedings of the ACL 2010 System Demonstrations*, 25–29. http://www.aclweb.org/anthology/P10-4005

Odijk, J. (2009). *Data categories and ISOCAT: Some remarks from a simple linguist* [Presentation given at FLaReNet/CLARIN Standards Workshop, Helsinki, 30 September]. https://surfdrive.surf.nl/files/index.php/s/ZoiwMlJ6mNKC9UP

Odijk, J. (2014). *Discovering resources in CLARIN: Problems and suggestions for solutions* [unpublished article, Utrecht University]. http://dspace.library.uu.nl/handle/1874/303788

Odijk, J. (2019). Discovering software resources in CLARIN. *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, (159), 121–132. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=159&Article_No=13

On the Successful Migration of Research Data

Claus Zinn and Thorsten Trippel
University of Tübingen, Germany

Abstract

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

1 Introduction

The maintenance of a research data repository comes with substantial costs. While a large part of the efforts is devoted to data curation, metadata annotation, and ingestion as well as the communication with data depositors and consumers, there is a significant workload involved in keeping the repository software up-to-date. Security updates are a major concern; at short notice, they must be played in a running repository system to make it less vulnerable to external threats. From time to time, old versions of repository systems are deprecated and stop benefiting from security patches. In this case, one has to perform a major upgrade to a newer version of the software. Costs related to software maintenance are rising, and some institutions may consider migrating their research data to an external, infrastructural organisation that is experienced with research data management (RDM), already hosts research data from a variety of other disciplines, and has trained staff. Scaling up certainly helps to keep expenditures in check. Ultimately, the turnaround of key personnel triggered our migration process.

2 Background

In 2010, our department started offering a repository system (called "Tübingen Archive for LAnguage Resources", short TALAR) to provide a centralised storage solution for all data created and maintained in our institution. We started with a system based upon Fedora Commons (https://fedora.lyrasis.org) (version 3), which we extended with a number of bells and whistles (*e.g.*, a GUI as well as a shell-based environment to support data ingestion and rights management; an OAI-PMH port to connect to the CLARIN data harvesting infrastructure). Due to security reasons, we later updated the system to version 4 of Fedora-Commons. Security patches available for this version where applied whenever possible.

Having control over your own repository comes with a significant amount of responsibility, but it also opens up a design space around many aspects of research data management, *e.g.*, how to best describe research data with metadata; and what research data should be accepted for ingestion (only internal data stemming from our own institution, or data from other institutions, or only data passing some quality threshold?). In the past decade, we built tools around the Fedora Commons repository system (Dima, Henrich, et al., 2012), the Bagman software for researchers to help them transferring their data to the archive (Zinn, 2022), the ProFormA editor to help them annotating their research data with metadata (Dima, Hoppermann, Hinrichs, Trippel, & Zinn, 2012), and we also created crosswalks between CMDI to Dublin Core and MARC-21 (Zinn, Trippel, Kaminski, & Dima, 2016).

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

3 The TALAR Repository

The TALAR repository has now been operational since 2010. In 2023, it contains over 600 datasets, totalling hundreds of gigabytes of data. Each dataset is richly described with CMDI-based metadata, and addressed by a handle-based persistent identifier. Usually, part identifiers are used to address individual files inside a dataset. Handles without part identifiers point to the HTML-based *landing page* of a dataset, which is automatically rendered by using a CMDI-2-HTML based transformation of the dataset's CMDI file. In TALAR, all research data is hierarchically structured. Moreover, the underlying Fedora-Commons software made is possible to assign access permissions at individual directory and file levels. Those *Access Control Lists* were used to give authenticated individual users read access to (non-public) datasets. TALAR is a certified CLARIN-B centre and takes part of the CLARIN harvesting infrastructure.

The migration of the entire repository created a situation were we examined all the datasets accumulated so-far at once, and this forced us to review the collections in their entirety. The bird's-eye perspective revealed that a significant part of the the SFB833 collection have datasets that consist of only a single text file (usually in PDF format). Those 265 PDF articles played the role of a reference collection to support scientific research across the many members of the CRC-833. While some of those papers have been published in journals, others were manuscripts submitted for review or were in press. Also, there is a small collection of thirteen datasets that should have gotten their own root node. Each dataset contains teaching material targeted at graduate and PhD students. Also, all course material (presentation slides, exercises, data files) come with an open CC-BY licence. Moreover, the repository also contains 54 service descriptions for the WebLicht workflow engine (Hinrichs, Zastrow, & Hinrichs, 2010). It is paramount that those CMDI-based WebLicht service descriptions must continue to be available.

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

Table 1: Overview of Target Repositories.

The bird's-eye view informed our decision to migrate our data to two different repositories using five different *communities*. Table 1 depicts the new organisation.

Goals for the Migration We found that migration must proceed along the following lines: (i) preserve the hierarchical structure of research data, but only if needed; (ii) migrate *all* research data; (iii) ensure that there is no information loss in terms of metadata; and (iv) strive for minimal service disruptions.

4 Migration

4.1 Migration to Zenodo

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

The ingestion of literature data was rather straightforward. With each PDF file being complemented with a CMDI-based metadata description, we wrote an XSLT stylesheet to extract the relevant metadata into the required DataCite fields, namely, author, title, publication date, and description. It showed that the CMDI files did not have more information that needed to be preserved, and therefore, no information loss incurred. Consequently, we did not ingest any CMDI files to the Zenodo CRC-833 community. Also, due to copyright issues, it was required that all research data in the Zenodo community "sfb-833-literature" was restricted. Interested parties can contact the community curator for which we have created a new special-purpose email account.The teaching material had a complex nature. They usually consisted of many, in part hierarchically structured, files using a variety of different data formats. We found their CMDI-based description rather shallow, not making use of the potential that the CMDI profile “CourseProfile” offered. As a result, we also omitted the ingestion of CMDI profiles to Zenodo. All teaching material in the Zenodo community “talar-teaching-material” comes with a CC-BY licence. In sum, 278 data sets left the realm of Tübingen University and found their new home in the Zenodo repository. All SFB-833 literature data was automatically ingested into Zenodo using a Python-based script that makes use of the Zenodo developer API. The teaching material was manually ingested into Zenodo, usually by creating zip archives to hold highly structured data.

4.2 Migration to FDAT, the Institutional Research Data Repository

There were 323 datasets still to be taken care of. In terms of content and size, they constitute the “real” research data. To mirror the high-level structure of the TALAR source repository, we have created three communities on the new institutional FDAT repository, see Tab. 1. The target repository defined a number of hard constraints that we needed to deal with: (i) CMDI-based metadata is not an accepted metadata standard; all research data must be described using DataCite; (ii) the size of each dataset is limited to 100 GB and cannot contain more than 100 individual files; (iii) access to a dataset is either restricted for all, or accessible for all. No individual rights can be associated with a dataset; and (iv) all datasets are assigned newly created, DOI-based persistent identifiers – Many of our datasets have a deep directory structure, which is not supported by FDAT. In these cases, the hierarchy was flattened, usually by replacing them with zip archives so that their unarchiving reestablishes the hierarchy. Moreover, some datasets had information duplicated. Sometimes, the files of a dataset were complemented with a ZIP archive that also held all files. Such redundancy was removed, consistently in preference for the ZIP archive. Hence, the migration required us to review each dataset individually. During the review, we observed other, mostly minor, oversights or flaws in the metadata, which we corrected in due course. We also contacted some of the researchers that produced the research data to review our migration work. This made us realize that an acceptable translation from CMDI to DataCite is far from trivial, and must take into account a few subtle but important aspects. – To avoid any loss of information given in the CMDI metadata description of a dataset, we made the CMDI file an integral part of the dataset itself. While metadata can be altered after the publication of a dataset, this is not the case for the research data itself, and consequently, it is crucial to ensure that all CMDI-based metadata is its final, publishable state. As a result, each CMDI file is diligently reviewed by the communities’ curator before the entire dataset it describes is published.

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

Persistent Identification The FDAT repository requires using DOI-based persistent identifiers. In principle, the FDAT repository allows the addressing of individual files of a dataset. For this purpose, however, no DOI can be used, only the resolved URL. However, the FDAT administrator cannot guarantee that the resolved URL will continue to be serviced in the future. If we were to continue supporting part identifiers, then we must continue to support our local resolver, and maintain its mapping table to do the rewriting, and change the rewriting whenever the target URL of FDAT changes. To minimize our commitments, we incrementally stopped our support for part identifiers; access to all research data is via their new FDAT-based landing page. All handles registered at `handle.net` will directly point to their respective DOI handles of FDAT, bypassing the local URL resolver, which can hence be retired. Handles used in CMDI files are replaced by their DOIs. Only users trying to invoke legacy handles with part identifiers will encounter a “404 – page not found” error.

Metadata Provision The migration required us to convert CMDI-based metadata to DataCite. To minimize information loss, some resource-specific metadata is written into one of Datacite’s description fields. Moreover, the original CMDI-based description becomes part of the research data stream so that users can consult the metadata for information that cannot be mapped to DataCite. – The source repository offered OAI-PMH harvesting for metadata in DC, MARC 21, and CMDI. Zenodo and FDAT, however, only support metadata harvesting for DataCite. Hence, we will continue to operate a designated OAI-PMH service that provides CMDI-based metadata to the Virtual Language Observatory (VLO).

Transitional period For new research data, our institution will continue to provide help to *local* researchers who would like to archive their research data in a trusted, sustainable environment. However, we will not accept any new research data from *external* parties. For the data we accept, end-users can expect to receive the same quality of service as before; annotation of research data with CMDI will continue. The CMDI will be part of the data set, and distributed via our OAIPMH service. Upon the closure of the institute, *all* users will be directed to another CLARIN-B centre, or to the FDAT archive manager.

5 Conclusion

The migration of research data is no easy matter and is bound to create issues that cannot always be resolved without making compromises. The actual effort for migrating all data involved numerous, internal discussions, coordination with the FDAT manager, the authoring of XSL-based stylesheets for ingest mechanisation, the adaptation of the WebLicht software to fetch their CMDI-based service descriptions in a modified manner, the editing of CMDI files, and the reconfiguration of handles to point to new target URLs. Our discussion shows that the migration of research data needs careful planning and execution, and that any migration efforts must be started well in advance. – The migration of research data freed us from maintaining a good number of software packages. ProForma, the Erdora shell and its GUI, the OAI-PMH plugin, the local resolver, and the entire Fedora Commons repository is to be retired. The only software that is run for a longer period of time is the OAI-PMH service to make available all CMDI files to the CLARIN VLO, ensuring that the visibility of our language resources stays high. – Our department has offered a repository system since 2010; it started at a time where research data management was underdeveloped in the infrastructural institutions of our university. We had little alternatives to perform research data management other than doing it ourselves. Supported by national and international funding from CLARIN, we developed an entire eco-system around RDM. This time is now coming to an end.

References

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

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

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

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

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

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

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

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

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

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

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

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

Cyprian Laskowski
University of Ljubljana
cyp@cjvt.si

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

Simon Krek
University of Ljubljana
simon.krek@ijs.si

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

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

1 Introduction

The Slovenian research infrastructure (RI) CLARIN.SI was founded in 2014 and became a member of CLARIN ERIC in 2015. So far, the only publication in the English language[1] that comprehensively presents CLARIN.SI was published shortly after its establishment (Erjavec et al., 2014), where we described the first steps of the RI and plans for further work. This paper summarises the state of the infrastructure ten years later.

2 Management of CLARIN.SI

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

CLARIN.SI is organised as a consortium with a current membership of 12 partner institutions.[2] The consortium brings together all the main institutions involved in the development or use of language resources and technologies in Slovenia, including all four Slovenian universities, four research institutes (including the Fran Ramovš Institute for the Slovenian Language of the Slovenian Academy), the National and University Library, two language technology companies, and the Slovenian Society for Language Technologies.

Decisions on the management of the RI are made or confirmed by the CLARIN.SI Management Board, in which each partner has one representative and one or more deputies. Communication takes place via the mailing list of the Board, which currently has 34 members, and at annual meetings where we discuss the work of the RI over the past year and make plans for the next.

CLARIN.SI maintains a bilingual (Slovenian, English) website[3] that presents the RI and its services.

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

3 The CLARIN.SI repository

The core service offered by CLARIN.SI is the maintenance of its repository of language resources and tools. The repository runs on the open CLARIN-DSpace platform,[4] which was developed within the

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

[2] https://www.clarin.si/info/partners/

[3] https://www.clarin.si/

[4] https://github.com/ufal/clarin-dspace

Czech CLARIN (originally under the name of LINDAT, now CLARIAH, after merging with the Czech DARIAH infrastructure) at the Institute for Formal and Applied Linguistics at Charles University in Prague.

The repository pages contain instructions for depositing entries[5] with particular attention given to the metadata required and how it should be formatted and to the acceptable encoding of deposited data.[6] In this respect, the CLARIN.SI repository differs from most other CLARIN repositories (Lenardič & Fišer, 2022), which typically only provide a list of formats without any instructions on how to describe and prepare the data. These additional instructions are particularly useful for humanities authors who may not have the necessary computer skills for data preparation. The repository currently contains over 600 entries, which are the result of the efforts of over 1,000 authors from 115 institutions.

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

4 Web services

In addition to the repository, CLARIN.SI maintains several other online services. The most significant by far are the concordancers, in particular noSketch Engine and KonText. Both use the same back-end program, namely Manatee (Rychlý, 2007), which enables fast querying of large and richly annotated corpora, but the two differ in their user interfaces. NoSketch Engine is an open-source version of the commercial Sketch Engine concordancer (Kilgarriff et al., 2014),[7] while KonText was developed at the Czech National Corpus Department of Charles University in Prague (Machálek, 2020). In order to facilitate the use of these concordancers by Slovenian users, both have been localised to Slovenian. Apart from their appearance, the main differences are that noSketch Engine offers a wider range of features (e.g. keyword extraction, better corpus statistics), while KonText supports log-in via the AAI system (the same as the repository), which then allows personalised screen settings, saving query history, etc. Recently, however, we installed two versions of noSketch Engine. One is intended for anonymous users, while the other supports (albeit self-registered) log-in and enables similar additional features as KonText.

In principle, all CLARIN.SI concordancers offer access to the same set of corpora, which currently comprise over 200 corpora in 41 languages, including reference, specialised, spoken, and multilingual parallel corpora. The CLARIN.SI concordances are employed in the study programmes at numerous universities, as part of linguistic research or in various research projects, as well as in Slovenian translation companies.

As part of the CLASSLA knowledge centre, which is discussed in the next section, CLARIN.SI also provides the CLASSLA Annotation Tool,[8] an online service for the automatic linguistic annotation of texts that uses the CLASSLA-Stanza language processing tool (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023) and offers models for the annotation of Bulgarian, Croatian, Macedonian, Serbian and Slovenian as well as models for non-standard (colloquial) Croatian, Serbian and Slovenian. It supports annotation on the level of lemmas, morphology, dependency syntax, named entities and, depending on the language, also semantic roles.

For controlled and collaborative development, the Git platform has become very popular, which we also use within CLARIN.SI, not only for software but also for language resources. On GitHub, CLARIN.SI has its virtual organisation,[9] which gathers over 100 open-source projects. Unlike GitHub, the GitLab platform can also be installed locally, and the GitLab installation on CLARIN.SI[10] contains around 20 projects, both public and private, the latter mostly to do with CLARIN.SI services.

[5] https://www.clarin.si/repository/xmlui/page/deposit

[6] https://www.clarin.si/repository/xmlui/page/data

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

[8] https://clarin.si/oznacevalnik/eng/

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

[10] https://gitlab.clarin.si/

5 Support and dissemination

Together with the Bulgarian CLARIN research infrastructure (CLADA-BG) and the Institute of Croatian Language, CLARIN.SI manages the CLARIN Knowledge Centre for South Slavic languages CLASSLA,[11] within the framework of which it provides assistance in the use of language resources and technologies for South Slavic languages. CLASSLA supports researchers with documentation on open language resources, tools for creating and processing text corpora, and other language technologies. In addition, the centre develops its own language technologies and corpora to meet the needs of South Slavic languages.

In 2021, CLARIN.SI also became a member of the CLARIN Knowledge Centre for Processing User-Mediated Communication CKCMC,[12] which is managed by Eurac Research, Bolzano.

CLARIN.SI provides financial support for projects selected through an annual call, which is open for members of the consortium. Since 2018, when the initiative started, 30 projects have been successfully implemented, producing, for example, the corpus of parliamentary debates of the National Assembly of the Republic of Slovenia siParl (Pančur et al., 2024), and the speech corpus Gos VideoLectures (Verdonik et al., 2019).

CLARIN.SI participates in the organisation of events in the field of computational linguistics and related topics in Slovenia, e.g. the 34th European Summer School in Logic, Language and Information (Ljubljana, 2023), and especially in the main conference for this field in Slovenia, i.e. the International Conference on Language Technologies and Digital Humanities, which takes place biennially in Ljubljana.

The operations of CLARIN.SI and its knowledge centres are regularly presented at national and foreign workshops and conferences, such as the ESFRI conference. CLARIN.SI also organises workshops on the use of corpora and language technologies. So, for example, we held workshops[13] for using the noSketch Engine concordancer, WebAnno and Git platforms, while the CLASSLA knowledge centre participated in a workshop on using corpora for analysis of the regional variation of gender marking in a language.[14]

Finally, we inform the public about the activities of the partners of the CLARIN.SI consortium and its knowledge centres through the latest news published on the website of the infrastructure, its mailing list and posts on the CLARIN.SI profiles on X, Discord, and LinkedIn.

6 Involvement in projects and infrastructures

CLARIN.SI participates in national and European projects, which promotes the utilisation of resources and improves visibility while simultaneously securing additional funding.

In the framework of the Slovenian 2018–2021 cohesion funds projects, three partners of the consortium upgraded their physical infrastructure to improve the speed and reliability of CLARIN.SI online services, while the GPU servers acquired by the University of Maribor serve for research into deep learning of language data processing.

Among the European projects, we highlight ELEXIS.[15] In order to meet the requirements of this project, a new collection was created in the CLARIN.SI repository, which contains metadata and links to web interfaces of 143 digital dictionaries. We also plan to establish a new CLARIN ELEXIS Knowledge Centre for Digital Lexicography within CLARIN.SI and JSI.

We have participated in a number of national projects. The largest to date was "Development of Slovenian in a Digital Environment",[16] where CLARIN.SI provided its services for reviewing and depositing language resources created within the project and the definition of schemas for the mark-up of Slovenian language resources.

[11] https://www.clarin.si/info/k-centre/

[12] https://cmc-corpora.org/ckcmc/

[13] https://www.clarin.si/info/dogodki/

[14] https://www.clarin.si/info/k-centre/workshops/

[15] https://elex.is/

[16] https://slovenscina.eu/en

CLARIN.SI cooperates with its Slovenian sister infrastructures from the field of social sciences (CESSDA-SI, i.e. ADP) and the humanities (DARIAH-SI). For example, in the project “RDA Node Slovenia” (2019–2020), coordinated by ADP (University of Ljubljana), we reviewed and analysed Slovenian research data repositories (Meden & Erjavec, 2021), while with DARIAH-SI (Institute of Contemporary History) we have a long history in the joint development of corpora of parliamentary data and schemas for their encoding.

CLARIN.SI is active in the work of CLARIN ERIC. We obtained funds for two smaller projects that included international workshops in 2016 (Ljubljana) and 2019 (Amersfoort). The latter, in cooperation with DARIAH-SI, was dedicated to the development of recommendations for the standardised coding of corpora of parliamentary debates under the name Parla-CLARIN[17] (Erjavec & Pančur, 2022), which has become a popular choice for encoding parliamentary corpora. On this basis, CLARIN.SI acquired a key role in two major “CLARIN Flagship” projects, ParlaMint I (2020–2021) and ParlaMint II (2022–2023).

The ParlaMint projects created comparable, interpretable and uniformly coded corpora of parliamentary debates. In ParlaMint I, CLARIN.SI led the collection and encoding of 17 corpora of national parliaments (Erjavec et al., 2022b), which are now openly accessible on the CLARIN.SI repository and on the concordancers. ParlaMint II expanded and enriched the existing corpora while also adding new ones, and resulted in the production of 29 corpora. CLARIN.SI members (co-)led four of the five work packages of the project.[18]

7 Conclusions

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

In the near future, we plan to implement tracking of the use of CLARIN.SI services, in particular the concordancers. This will not only provide us with some key performance indicators, but also help us to focus on the implementations and corpora that are used the most in order to concentrate development where it will have the greatest impact.

The main long-term challenge for CLARIN.SI is to increase activities in the areas of education, training and support in the use of the infrastructure for existing and future users. This includes training activities, including “training the trainers”, extended online documentation and tutorials, and other outreach activities. A step in this direction is the currently on-going (April to September 2024) CLASSLA-Express series of tutorials, which comprise hands-on exercises on using the CLASSLA web corpora in the CLARIN.SI concordancers.[19] The six CLASSLA-Express workshops take place in Croatia (Zagreb and Rijeka), Serbia (Belgrade), North Macedonia (Skopje), Bulgaria (Sofia) and Slovenia (Ljubljana).

In addition, there is also a need to strengthen support activities – for example, by providing assistance in creating a data management plan for students and researchers in the humanities and by integrating corpora submitted to the repository into the concordancers.

Acknowledgements

We would like to thank the three anonymous reviewers for their helpful suggestions. The presented work was supported by the Slovenian research agency ARIS in the scope of its funding for ESFRI research infrastructures, as well as many other projects mentioned in the paper.

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

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

[19] https://www.clarin.si/info/k-centre/workshops/classla-express/

References

Erjavec, T., Dobrovoljc, K., Fišer, D., Javoršek, J. J., Krek, S., Kuzman, T., Laskowski, C. A., Ljubešić, N., & Meden, K. (2022a). Raziskovalna infrastruktura CLARIN.SI (The CLARIN.SI research infrastructure). *Proceedings of the Conference on Language Technologies and Digital Humanities*, 47–54. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Erjavec-et-al_Raziskovalna-infrastruktura-CLARIN.SI.pdf

Erjavec, T., Javoršek, J. J., & Krek, S. (2014). Raziskovalna infrastruktura CLARIN.SI [https://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf]. *Zbornik Devete konference JEZIKOVNE TEHNOLOGIJE*. https://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf

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

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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

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

Ljubešić, N., & Dobrovoljc, K. (2019). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 29–34. https://doi.org/10.18653/v1/W19-3704

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface [https://www.aclweb.org/anthology/2020.lrec-1.865]. *Proceedings of the 12th Language Resources and Evaluation Conference*, 7003–7008. https://www.aclweb.org/anthology/2020.lrec-1.865

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

Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2024). Slovenian parliamentary corpus (1990-2022) siParl 4.0 [Institute of Contemporary History]. http://hdl.handle.net/11356/1936

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, 65–70.

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.

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

A CLARIN Resource Family for Corpora of Communication Disorders

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

Nicola Bessell
University College Cork, Ireland
n.bessell@ucc.ie

Katarzyna Klessa
Adam Mickiewicz University, Poznan, Poland
klessa@amu.edu.pl

Alice Lee
University College Cork, Ireland
a.lee@ucc.ie

Satu Saalasti
University of Eastern Finland, Finland
satu.saalasti@uef.fi

Eric Sanders
CLS, Radboud University, Netherlands
eric.sanders@ru.nl

Abstract

This paper describes a new branch for the CLARIN Resource Family: namely one for corpora with speech from individuals with language and speech disorders (CSD): https://www.clarin.eu/resource-families/corpora-disordered-speech. We explain why the initiative was taken by the DELAD steering group, followed by our method for collecting and detecting relevant resources, and the initial results. We also introduce a new CMDI (Component Metadata Infrastructure) profile for such corpora with components for speech sound disorders and language disorders that we recommend for (new) corpora of this resource family.

1 Introduction

In 2018 Fišer, Lenardič and Erjavec (2018) presented the CLARIN Resource Family initiative with the goal "to collect and present in a uniform way the most prominent data types in the network of CLARIN consortia that display a high degree of maturity, are available for most EU languages, are a rich source of social and cultural data, and as such are highly relevant for research from a wide range of disciplines and methodological approaches in the Digital Humanities and Social Sciences as well as for cross-disciplinary and trans-national comparative research". This initiative has proven very successful. Meanwhile the webpage on CLARIN Resource Families (https://www.clarin.eu/resource-families) lists 26 members consisting of 15 corpora families, 6 lexical resource families and 5 tool families. A resource family that was missing was one for corpora with speech from individuals with language and speech disorders (CSD).

CSD are invaluable resources for education and research. However, they are costly and hard to build and can be difficult to share given various issues, such as the preservation of privacy and confidentiality of the participants, and the possible extra work and cost required for formatting the datasets for comparable sharing and hosting in a repository. Overcoming these challenges is important, as sharing data enables better science in the future. Re-analysis of raw data fosters improvement in the reproducibility and robustness of research. The availability of datasets allows other research teams to answer a different research question which maximizes the value of the data collected and in turn increases the impact of research of the original investigators. The availability of data also facilitates systematic review and meta-analyses. Datasets that are comparable can be pooled together to form a bigger set of data permitting more sophisticated analyses. The pooling of similar datasets also allows cross-linguistic research between countries, or investigation of rare conditions as it is often difficult to collect data from a sufficient number of participants by a single research center (see Lee et al., 2022). Hence, it greatly benefits the discipline if more researchers in the area of clinical linguistics and phonetics, and speech and language therapy (or speech-language pathology) consider sharing speech data. This CLARIN Resource Family is designed to exactly serve this objective.

The authors of this paper considered it as part of their expertise and involvement with CLARIN to propose this resource family and implement it. Upon approval of our proposal by the CLARIN board the initiative started in the fall of 2023, and was completed in August 2024.
All authors (except Sanders) are members of the steering group of DELAD, and in the next section we will explain what DELAD is and explain its relationship with CLARIN via its K Centre for Atypical Communication Expertise (ACE). Then we will address the methods for collecting the resources for the resource family and its results (section 3). In section 4 we will introduce a CMDI (Component Metadata Infrastructure) profile for Corpora of Communication Disorders which includes metadata components and elements for specific disorders. We conclude (section 5) with the activities accomplished so far and future work.

2 The DELAD initiative

DELAD stands for Database Enterprise for Language And speech Disorders. The acronym is also a word in Swedish, meaning “shared”. The initiative, previously known as DisorderedSPeechBank (Ball et al., 2016), was started by Professors Nicole Müller and Martin Ball in 2015 when they were with Linköping University, Sweden (Lee et al., 2023; Lee et al., 2022). The present aim of DELAD is to facilitate the sharing of corpora of speech of individuals with communication disorders (CSD) among researchers and educators in an ethical and secured manner.

The work of DELAD has been done mainly through workshops and regular meetings of the steering group. Six workshops, each with specific themes in the broader context of ethics and infrastructures for CSD sharing, have been held (Lee et al., 2023; Lee et al., 2022; see also the Workshops section on the DELAD website: http://delad.net). Researchers or specialists with pertinent expertise including speech and language pathology, infrastructure for data archiving, intellectual property rights, ethics and the General Data Protection Regulation / GDPR, have been invited to present and discuss relevant topics at the workshops. The first two in Linköping were funded by Riksbankens Jubileumsfond and the other four by CLARIN (Common Language Resources and Technology Infrastructure). The seventh workshop, again supported by CLARIN, is taking place on 11-12 September 2024 as an associated event with the 21st International Congress of Linguists (ICL), Poznan, Poland.

Other key achievements of DELAD include refurbishing the website (see the URL above) with information and an application form for connecting with the initiative, linking with the CLARIN K-Centre for Atypical Communication Expertise (https://ace.ruhosting.nl/) for hosting and accessing CSD through two CLARIN B-Centres: The Language Archive (https://archive.mpi.nl/tla/) and TalkBank (https://talkbank.org/) (Lee et al., 2022); information on annotation tools and techniques for disordered speech, guidelines on consent and data storage, and a role playing activity for learning Data Protection Impact Assessment (DPIA) (Lee et al., 2023). The last three resources can be accessed via the DELAD website.

3 Methods and preliminary results for collecting resources

In order to identify potential resources for the new Catalogue of Corpora of Disordered Speech we have both reviewed the existing CLARIN resources that fall into the category. We made an inventory of the material (datasets and resources) offered through DELAD and CLARIN centres with expertise in CSD. For DELAD we departed from https://delad.ruhosting.nl/wordpress/data-inventory/ and consulted our members for any updates. For Talkbank we concentrated on the relevant resources in Talkbanks Clinical Banks (https://talkbank.org/). Further, we inspected any other relevant datasets in the CLARIN’s Virtual

Language Observatory (VLO), the ELRA catalogue and ELRA's LRE Map. Finally, we inspected potential candidates to be found in other resource families.

We prepared an online questionnaire using Google Forms and asked potential contributors to submit information regarding each dataset that could be registered as a resource. We included details that were required for the CLARIN resource listing. These were details regarding the name, corpus URL (if available), description, speaker and disorder characteristics, language and size of the dataset and possible annotations. Furthermore, contributors were asked to identify any possible publication related to the dataset and, also, whether the data are licenced. The questionnaire was distributed via email for previous DELAD workshop attendees and other researchers that had previously indicated interest in sharing their datasets. The name and email submitted via the questionnaire were solely used for the purpose stated in the questionnaire. At a later stage the questionnaire was also submitted to a much wider audience via the CLARIN distribution channels (social media, newsflash, UIC mailing list, corpora mailing list).

As an overview, the resulting resources included both new and already identified corpora, for a total of 30 corpora. The search of CLARIN resources identified 9 relevant corpora in existing CLARIN repositories and 7 collections from TalkBank. Based on the existing DELAD inventory and a survey of DELAD members and affiliates we identified 12 corpora, some of which are also included in TalkBank or CLARIN repositories. Finally, there are 11 additional corpora in other repositories. A broad range of speech sound disorders as well as language disorders are represented in the corpora. A range of language families are represented, although English is predominant. Further, the datasets cover e.g. samples from speech sound disorders in children or parallel datasets from hearing and hearing-impaired adults. The modalities of the datasets may include varied types of audio, video and acoustic-articulatory data.

4 A CMDI profile for Corpora of Communication Disorders

In order to adhere to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) as well as possible, the metadata of the DELAD corpora will be included in the CLARIN Virtual Language Observatory (VLO) (Van Uytvanck et al., 2012). To this end, the metadata will be stored in CMDI format. CMDI is an interoperable format to share metadata which is used in CLARIN's Virtual Language Observatory to make datasets more findable and (its metadata) more accessible. We created an adapted version of the CorpusCollection profile (identifier: clarin.eu:cr1:p_1493735943947) by adding a few newly created components concerning language and speech disorders. The identifier of the new CSD profile is: clarin.eu:cr1:p_1708423613606. The CorpusCollection profile was developed to store metadata in the Corpus Collection, a collection of metadata of various resources at the Radboud University. The metadata in the Corpus Collection is harvested by the VLO.

The new profile contains two added components. Both components contain a number of elements and one component consisting of elements that are associated with the corresponding disorder. All elements are optional and of the Boolean type, i.e. of each disorder can be indicated whether they do or do not appear in the corpus, or it can be left out. Below is an overview of the elements of both components.

Speech sound disorders *(Component under Resource)*

- Developmental dysarthria *(Element)*
- Developmental verbal dyspraxia / Childhood apraxia of speech *(Element)*
- Dysarthrias (in adults) *(Element)*
- Apraxia of speech (in adults) *(Element)*
- Fluency disorders (including stuttering and cluttering) *(Element)*
- Voice disorders *(Element)*
- Phonological disorder *(Element)*
- Speech sound disorders associated with *(Component under Speech sound disorders)*
    - Hearing loss and/or cochlear implants *(Element)*
    - Genetic syndromes *(Element)*
    - Autism *(Element)*
    - Aphasia *(Element)*
    - Cleft palate and velopharyngeal dysfunction *(Element)*
    - Head and neck cancer *(Element)*

    - Cognitive disorders (including traumatic brain injury, dementia) *(Element)*
    - Psychiatric disorders *(Element)*
    - Neuropsychiatric disorders (e.g. ADHD, Tourette syndrome) *(Element)*

Language disorders and delay *(Component under Resource)*

- Language delay *(Element)*
- Developmental language disorder *(Element)*
- Aphasia *(Element)*
- Language disorders associated with *(Component under Language disorders and delay)*
    - Hearing loss and/or cochlear implants *(Element)*
    - Intellectual disabilities *(Element)*
    - Genetic syndromes *(Element)*
    - Autism *(Element)*
    - Cleft palate *(Element)*
    - Cognitive disorders (including right hemisphere damage, traumatic brain injury, dementia) *(Element)*
    - Psychiatric disorders *(Element)*
    - Neuropsychiatric disorders (e.g. ADHD, Tourette syndrome) *(Element)*

5 Conclusion and future work

The current project has been very useful to identify, and continues identifying, the most prominent corpora related to disordered speech available in a variety of EU languages. These datasets are a rich source of social and cultural data, and listing them will improve their findability and increase their research use in a variety of methodological approaches. The resulting new resource family page is published under https://www.clarin.eu/resource-families/ as https://www.clarin.eu/resource-families/corpora-disordered-speech.

References

Ball, M., Baqué, J., Beck, J., Beijer, L., Bernhardt, M., Bressmann, T., Hertrich, I., Howard, S., Klippi, A., Kristoffersen, K. E., Lee, A., Mildner, V., van den Heuvel, H., Müller, N., Lundeborg Hammarström, I., & Samuelsson, C. (2016, June). *The DisorderedSPeechBank: A multilingual digital archive of disordered speech* [Paper presentation]. The 16th Meeting of the International Clinical Phonetics and Linguistics Association, Halifax, Canada.

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)*, http://www.lrec-conf.org/proceedings/lrec2018/pdf/829.pdf

Lee, A., Bessell, N., van den Heuvel, H., Klessa, K., & Saalasti, S. (2023). The DELAD initiative for sharing language resources on speech disorders. *Journal of Language Resources and Evaluation*, https://doi.org/10.1007/s10579-023-09655-2

Lee, A., Bessell, N., van den Heuvel, H., Saalasti, S., Klessa, K., Müller, N., & Ball, M. J. (2022). The latest development of the DELAD project for sharing corpora of speech disorders. *Clinical Linguistics & Phonetics, 36*(2-3), 102-110. https://doi.org/10.1080/02699206.2021.1913514

Van Uytvanck, D., Stehouwer, H., & Lampen, L (2012). Semantic metadata mapping in practice: the virtual language observatory. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*. pp. 1029-1034. https://pure.mpg.de/rest/items/item_1454694_11/component/file_1478393/content

A Development Outlook for CLARIN's Northernmost Center

Steven Coats
University of Oulu, Finland
steven.coats@oulu.fi

Abstract

This paper introduces a new CLARIN-D center in Oulu, Finland, and discusses plans for the development of the center's resources in the context of coordination with Fin-CLARIN and the larger CLARIN community. Currently, the center hosts CoANZSE Audio, a searchable corpus of Australian and New Zealand speech transcripts and recordings. In coming years, it is hoped that the center can develop its online presence and offerings, focusing on structured, annotated, and curated computer-mediated communication data, especially multimedia data comprising text, audio, and video. The paper discusses the shareability of such data in the context of legislation permitting research use of copyrighted materials. The CoANZSE Audio resource is briefly introduced.

1 Introduction

In the autumn of 2023, a CLARIN-D center was established in Oulu, Finland, to provide access to a searchable audio corpus of Australian and New Zealand English, CoANZSE Audio. The initiators of the new center, whose creation was facilitated by the expertise of Netherlands-based CLARIN staff, subsequently recognized the need to make more structured multimedia resources available to the research community at large, for example annotated audio, video and computer-mediated communication (CMC) content. In recent years CMC has increasingly shifted towards interactive modalities that combine text, audio, video, and graphical content, often via services provided by commercial platforms. This development, facilitated by increases in processing power, available bandwidth, and storage capacities, as well as standardization of protocols for sharing of multimedia data, has resulted in complex interaction environments which often feature multimedia streams in addition to textual content. Curated corpora of multimedia data that can serve as the basis for research in language and interaction, however, remain few.

In this paper plans for developing the Oulu CLARIN center are described. First, related language resource infrastructures and CLARIN endeavors are briefly introduced in order to highlight the ecological niche of the Oulu hub: FIN-CLARIAH,[1] Finland's principal language resources and social science and humanities research infrastructure, and CKCMC, the CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media,[2] whose remit has significant synergies with the planned Oulu focus on multimedia content in English and other languages. The legal contexts for the collection and reuse of social media and other online data under copyright are then discussed, and the current content of the Oulu hub, CoANZSE Audio,[3] is introduced. The text closes with an outlook towards development of the Oulu CLARIN center towards C or B status.

[1] www.fin-clariah.fi

[2] https://cmc-corpora.org/ckcmc

[3] CoANZSE Audio

2 Related Language Resource Infrastructure

Finland is known for the high quality of its computational research infrastructure, which facilitates institutional initiatives not only in natural sciences and engineering, but also in corpus linguistics, natural language processing, and related fields. Finland's main research infrastructure for social sciences and humanities is FIN-CLARIAH, comprising FIN-CLARIN and DARIAH-FI; the infrastructure hosts the Language Bank of Finland,[4] which provides access to a wide range of language resources, tools, and services, with a focus on resources related to Finland's national languages of Finnish and Swedish, minority languages with official status in Finland (the Northern, Inari, and Skolt Sámi languages), as well as Finno-Ugric languages in general. As of early 2024, some multimedia content is available via the Language Bank, but social-media-platform-sourced multimedia content is not a primary focus. While it provides access to some English-language resources, its mission is primarily to provide access to resources in the national languages which have been created by Finnish researchers. In the context of increasing internationalization and the global ubiquity of social media content, Oulu's new CLARIN center is conceived as an infrastructure that will host resources in any language and created by researchers with any background, with a focus on social-media-sourced multimedia content.

The Oulu CLARIN center will benefit from the collective expertise of the existing CLARIN K-center for Computer-Mediated Communication and Social Media Corpora (CKCMC), which provides information to researchers and students interested in CMC-related resources and technologies. CKCMC is tasked with supporting members and interested parties in the production, modification, and publication of relevant resources and technologies, as well as organizing training activities. The Oulu center has strong links to CKCMC, and the related annual CMC-Corpora conference will be a suitable venue for research conducted on the basis of Oulu content, as well as workshops and other activities focused on the creation of the resource infrastructure.

3 Oulu CLARIN Center

3.1 Legal Contexts

Oulu's CLARIN hub plans to provide limited access to materials harvested from commercial platforms. The motivation to develop Oulu's CLARIN center has been prompted by recent changes in EU law which mitigate some copyright concerns, as well as by precedents and case law from the US and other Anglophone countries.

In the United States, where many large social media and streaming platforms are based, the use of copyrighted materials for purposes of research or scholarship is generally permitted according to the "Fair Use" provisions of copyright law, a condition that also holds for the wider Anglosphere in the UK, Australia, and New Zealand, where the exception is known as "Fair Dealing". U.S. Code Title 17, § 107,[5] stipulates that the legality of copying and utilizing copyrighted material is to be considered according to the purpose and character of the use, the nature of the copyrighted work, the amount and substantiality of the portion used in relation to the copyrighted work as a whole, and the effect of the use upon the potential market for or value of the copyrighted work. US law states that "reproduction in copies or phonorecords or by any other means specified by that section for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright" (USC 17 Section 107, n.d.). The analogous legislation in the UK can be found in sections 29 and 30 of the Copyright, Designs and Patents Act of 1988 (UK Copyright, Designs, and Patents Act, 1988), which licenses re-use and copying of materials under copyright for "research and private study" and "criticism or review".

In the EU, a 2019 directive (EU, 2019) was introduced to update 2001 legislation pertaining to copyright law and sharing of data (EU, 2001). The update explicitly authorizes text and data mining of material under copyright for purposes of research and education. Although the law does not discuss sharing of mined data, the issue is addressed by Recital 15, which notes that "uses for the purpose of scientific

[4] https://kielipankki.fi

[5] See also https://fairuse.stanford.edu/overview/fair-use

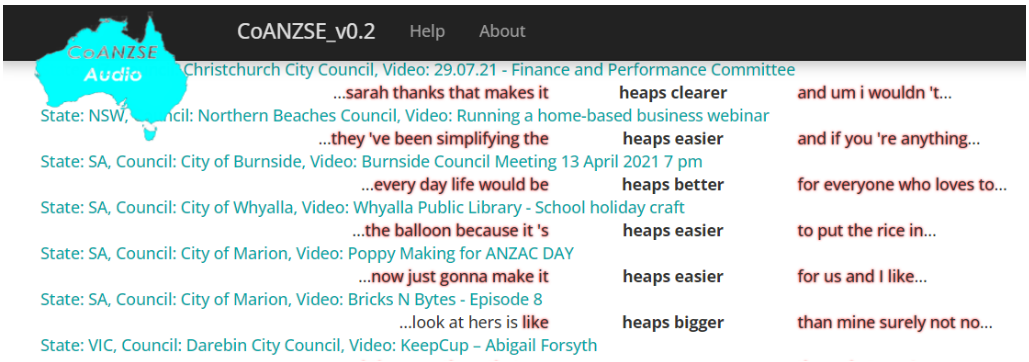
research, other than text and data mining, such as scientific peer review and joint research, should remain covered”, further referring to exceptions and limitations licensed by Article 5(3) of the earlier directive. That legislation notes exceptions including “use for the sole purpose of illustration for teaching or scientific research” as the first listed item. Since 2019, most EU member states have implemented EU 2019/790 as national legislation; this includes Finland in 2023.

The provisions of the EU’s AI Act, passed in June 2024 (EU, 2024), while not directly relevant to the activities of the planned Oulu center, nevertheless support an interpretation by which mining and reuse of social media content for non-profit research and educational purposes is permissible. Specifically, Recital 109 notes that compliance “should be commensurate and proportionate to the type of model provider, excluding the need for compliance for persons who develop or use models for non-professional or scientific research purposes”.[6]

These legal frameworks suggest that sharing for research purposes of data obtained from copyrighted content, for example as searchable extracts, is permissible use. Such an interpretation is in line with the general trend in academic research towards making research data available to the research community as searchable online resources, in line with FAIR principles (findability, accessibility, interoperability, and reusability; Wilkinson et al., 2016). In order to satisfy legal requirements, data hosted at Oulu’s CLARIN hub will not be provided in the same formats as on the original platforms, and will be removed at the request of rights holders.

3.2 CoANZSE Audio

Currently, Oulu’s CLARIN hub provides access to CoANZSE Audio (Coats, 2024), a searchable online corpus of more than 20,000 hours audio from Australia and New Zealand. The resource, created from videos uploaded to the YouTube channels of councils and other local government entities, updates the existing Corpus of Australian and New Zealand Spoken English (Coats, 2022a, 2022b) with audio and forced alignment data in addition to the textual content of the ASR transcripts. The search interface for CoANZSE Audio is a customized implementation of BlackLab (de Does et al., 2017), based on Apache Lucene and developed at the Dutch Language Institute.[7].



Christchurch City Council, Video: 29.07.21 - Finance and Performance Committee
...sarah thanks that makes it heaps clearer and um i wouldn 't...
State: NSW, cil: Northern Beaches Council, Video: Running a home-based business webinar
...they 've been simplifying the heaps easier and if you 're anything...
State: SA, Council: City of Burnside, Video: Burnside Council Meeting 13 April 2021 7 pm
...every day life would be heaps better for everyone who loves to...
State: SA, Council: City of Whyalla, Video: Whyalla Public Library - School holiday craft
...the balloon because it 's heaps easier to put the rice in...
State: SA, Council: City of Marion, Video: Poppy Making for ANZAC DAY
...now just gonna make it heaps easier for us and I like...
State: SA, Council: City of Marion, Video: Bricks N Bytes - Episode 8
...look at hers is like heaps bigger than mine surely not no...
State: VIC, Council: Darebin City Council, Video: KeepCup – Abigail Forsyth

Figure 1: Excerpt from CoANZSE Audio search results

Figure 1 shows a screenshot from the search interface: the results of a search for *heaps* followed by an adverb. CoANZSE Audio is accessible via Shibboleth federated login to members of the Edugain consortium, comprising universities and educational institutions worldwide, and is one of only a handful of large, searchable corpora of transcripts, audio, and alignments of English that are freely available for research and education (cf., e.g., Ljubešić et al., 2022).

As of August 2024, the site currently averages approximately 9 visits per day. Data from CoANZSE

[6] Other recitals, however, explicitly refer to the rights of copyright holders pursuant to Article 4(3) of Directive (EU) 2019/790; this suggests that while research use of social media content is permissible, compliance with any removal requests submitted by rights holders is required.

[7] http://inl.github.io/BlackLab

Audio and the underlying indexed transcript files have been used in studies of politeness phenomena (Thaler & Elsweiler, 2023) and of syntactic features (Morin & Coats, 2023), and a large scale study using corpus phonetic methods is in progress.

4 Summary and Outlook

Oulu's new CLARIN hub has been established in order to provide access to a speech corpus. In coming years, it is hoped that the hub will grow and will be able to host a variety of language resources, especially multimedia resources from social media platforms, including YouTube, Twitch, Tiktok, and other popular platforms, and potentially including streams from virtual reality platforms. Pending availability of funding, research data, and personnel support, Oulu's hub will expand and seek certification as a C or B CLARIN center, providing a modern an internationally oriented platform for researchers worldwide looking to analyze online multimedia data, and further cementing Finland's status as a leading country for the computational analysis of social sciences and humanities data.

References

Coats, S. (2022a). The Corpus of Australian and New Zealand Spoken English: A new resource of naturalistic speech transcripts. *Proceedings of the The 20th Annual Workshop of the Australasian Language Technology Association*, 1–5. https://aclanthology.org/2022.alta-1.1/

Coats, S. (2022b). *The Corpus of Australian and New Zealand Spoken English*. https://cc.oulu.fi/~scoats/coanzse.html

Coats, S. (2024, May). CoANZSE audio: Creation of an online corpus for linguistic and phonetic analysis of Australian and New Zealand englishes. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 3407–3412). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.302

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. *CLARIN in the Low Countries*, 245–257.

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2001). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A32001L0029

Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (2019). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689&qid=1724915900716

Ljubešić, N., Koržinek, D., Rupnik, P., & Jazbec, I.-P. (2022, June). ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the workshop parlaclarin iii within the 13th language resources and evaluation conference* (pp. 111–116). European Language Resources Association. https://aclanthology.org/2022.parlaclarin-1.16

Morin, C., & Coats, S. (2023). Double modals in Australian and New Zealand English. *World Englishes*. https://doi.org/10.1111/weng.12639

Thaler, M., & Elsweiler, C. (2023). The role of gender in the realisation of apologies in local council meetings: A variational pragmatic approach in British and New Zealand English. *Zeitschrift für Anglistik und Amerikanistik*, *71*(3), 217–239.

U.S. Code Title 17 - Copyrights Fair Use (n.d.). https://www.govinfo.gov/content/pkg/USCODE-2022-title17/html/USCODE-2022-title17-chap1-sec107.htm

UK Copyright, Designs, and Patents Act (1988). https://www.legislation.gov.uk/ukpga/1988/48/contents

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ..., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9.

Text collections as data at the National Library of Latvia

Anda Baklāne
Dep. of Digital Development
National Library of Latvia,
Latvia
anda.baklane@lnb.lv

Abstract

In recent years, the National Library of Latvia (NLL) has focused on making digital collections and bibliographical data accessible as datasets. Among the available offerings, textual datasets, particularly text corpora, stand out as the most prominent. These include the LatSenRom, a comprehensive corpus of Latvian Early novels, and extensive collections of historical daily newspapers available in plain text, lemmatized formats, and as embeddings. The design of these data sets is tailored to the specific aims and methodologies of individual research projects, allowing for considerable flexibility. The presentation highlights the approaches employed in developing these significant textual datasets, and explores their design and some use applications for research and analysis, underscoring their importance and versatility in academic and professional contexts.

1 Introduction

In addition to providing traditional databases and digitized collections to users, NLL offers a suite of digital research services tailored to meet the needs of digital scholarship. These services emphasize the use of the library's rich collections as data, encompassing text, images, and bibliographical data.

A crucial aspect of the services offered is the provision of datasets on demand. Researchers can request the creation of customized datasets based on specific criteria such as keywords, genres, time periods, or other parameters.

Some of these datasets are meticulously curated, involving a complex selection process, the enrichment of metadata, and the correction of OCR errors to ensure high-quality resources. The design of data sets, although based on some general principles, is dependent on the aims and methodologies of a particular study, hence there is more than one way for designing sound datasets. Text corpora are made available in plain text format, processed with LV-PIPE annotations, or as embeddings.

One of the challenges related to amassing corpora spanning long time periods is the change in writing tradition and grammar. At the beginning of the 20th century, Latvian publishing gradually transitioned from the old Gothic (Fraktur) to the Antiqua typeface, and the rules of orthography were changing as well. It was important for the library to address the needs of different types of researchers: those interested in studying transcriptions that are as close as possible to the original texts, and those interested in the comparative analysis of works from different periods and, hence, required normalized (modernized) versions of texts.

The datasets are also accessed online through the dedicated portal korpuss.lnb.lv, Latvian National Corpora Collection korpuss.lv, and in the CLARIN repository. Additionally, the library provides access to a specialized Jupyter Notebook environment via lab.lnb.lv. This platform allows researchers to perform data analysis within a robust and flexible computing environment.

The NLL is committed to supporting users throughout their research journey. It fosters active collaboration between the library staff and researchers to optimize the data curation process.

2 Latvian Early Novels

The corpus of Latvian Early Novels (after this - LatSenRom) includes novels written in Latvian and published in book format from 1879 to 1940. The foundation of the corpus is based on the "Latvian Novel Index" compiled by R. Briedis and A. Rožkalne in 2014.

The creation of the oldest parts of the corpus began as part of an international project - the COST Action 'Distant Reading for European Literary History' (CA16204), which took place from 2017 to 2022. The main task of the project was the creation of the "European Literary Text Collections" (ELTeC), a collection of data sets containing novel corpora in various languages, developed according to uniform principles. Each corpus was to include 100 works, published in European countries from 1840 to 1920, and carefully selected from the total number of publications according to balancing criteria: a proportional number of works from each decade, categories of author's gender, work length, and canonical status; the selection algorithm also stipulates that no more than three works from any single author are to be included in the corpus.

This experience led the creators of LatSenRom to adopt a different approach to its further development. Instead of creating a representative sample, they decided to include all works that are original novels in Latvian, published in book form either within the current territory of Latvia or beyond. This approach raises new questions, including the literary-theoretical challenge of determining which texts are considered novels. Practical issues are also relevant, such as why to choose a monograph instead of a periodical publication as the first edition, and how to treat separate units in cases of duologies, trilogies, and tetralogies. Some choices are dictated by practical considerations; for example, it is more practical to divide trilogies into parts, as they are often initially published separately (the date of the first publication is considered). Therefore, it must be acknowledged that although the data set as a whole is intended as a complete collection of first editions, some deviations are allowed in practice, and the boundaries of the data set are not absolutely strict: over time, some works may be excluded from the collection, while others may be added as the previously undiscovered works are found or ineligible works removed.

Between 1879 and 1940, publishing in Latvia transitioned from the Fraktur typeface to Antiqua, accompanied by changes in orthographic rules. Consequently, to facilitate comparative analysis of the literary works in the corpus and to utilize natural language processing tools designed for modern language, normalization (or modernization) of both the script and orthography was necessary.

3 Historical periodicals

The collection of digitized newspapers includes over 80% of all periodicals published before 1990, along with significant portions of modern digitized and digitally created newspapers and magazines. These materials have been segmented and subjected to optical character recognition, enabling users to utilize the capabilities of full-text search.

Due to the reservations related to access to copyright-protected materials, the NLL currently does not provide a public API for retrieving the materials from its collections. To facilitate access to most widely used resources, several titles of large 19th and 20th-century daily newspapers and magazines (e.g., Cīņa, Literatūra un Māksla, Karogs, Padomju Jaunatne) were made available as corpora on the korpuss.lnb.lv (former nosketch.lnb.lv) platform that allows users to access the statistics of word use and browse concordances without providing access to the full text of articles (the access to the datasets containing full text is still provided for research purposes).

Over time, the library began receiving an increasing number of requests from researchers for materials related to specific topics or research areas. In some cases, the coverage of these topics spanned several decades and included many periodical titles. Consequently, topic-based curated datasets have emerged as the second most popular type of dataset provided for research. Handling preselected sets of materials often proves more practical, as it reduces size and facilitates subsequent analysis steps in the research process.

Although there are plans to develop a selection tool that would allow researchers to autonomously compile a corpus without the aid of a digital librarian, certain use cases demonstrate that keyword-based selection alone often fails to exclude irrelevant results. Therefore, several steps of data collection are necessary to ensure that the corpus contains materials relevant not only to the keyword but also to the context and genre.

The normalization of historical texts is also highly relevant for periodicals. Unlike LatSenRom, which contains the oldest texts dating from the late 19th century, historical periodicals that span from the 18th century onward require a more complex approach to normalization. This involves several stages of gradual orthographic changes.

4 Versions and iterations of corpora

The NLL offers several variations of datasets, including changes over time (such as adding missing data and correcting errors), original and normalized versions, and versions that undergo different types of processing or annotation (such as morphological tagging, dependency parsing, NER annotation, and various types of embeddings). Managing these variations presents challenges in version control, naming, and organization of the datasets. The concept of iteration is employed to differentiate between parallel versions of texts and those that evolve over time through corrections, additions, or removals of works. Iterations of the same original object are datasets that feature different levels of normalization, mark-up, or annotation but are not considered less or more complete or accurate relative to one another.

In the case of the texts in part or fully published in Fraktur script, the normalization of the corpus is realized in two steps: normalizing of the script (typeface) and normalizing of grammar. The second step is yet not fully implemented in the time of writing this paper.

In addition to the technical challenges associated with developing algorithms for normalization, it is often unclear which aspects of normalization are relevant to a specific dataset or use case. For example, some texts may be written in a dialect of the Latvian language, or an author might have employed a unique style of speech for a character or in prose fiction. Normalizing these texts could erase differences that are potentially significant for text analysis.

Finally, to accommodate the needs of various research projects, the data is also processed using a range of natural language processing tools. As demonstrated in previous projects, such as the COST Action 'Distant Reading for European Literary History', layering annotations may not be always advisable. This approach can lead to conflicts between different annotation schemes and make the corpus version more challenging to access for researchers who lack the skills to work with complex structured datasets.

5 Conclusion

The National Library of Latvia has accumulated extensive experience in creating text-based datasets for researchers. Significant advancements have been made in making collections and data more accessible to researchers, offering support in designing and preprocessing datasets, and in developing versions and iterations tailored to specific research projects and use cases.

There remains significant work to enhance the version control of datasets and advance toward more sophisticated corpus design models, particularly in the case of carefully curated corpora like LatSenRom or topic-based corpora of historical periodicals. Steps must be taken to publish this data in a manner that is both open and secure, adhering to the principles of open and FAIR data while also complying with copyright restrictions. Additionally, numerous details require clarification concerning which datasets need long-term preservation and which may be considered temporary and can be discarded once they have served their immediate purpose.

References

Baklāne, A., Saulespurēns, V., Ozols, A., Krasovska, M. (2021). *Corpus of Latvian Early Novels*. CLARIN-LV digital library at IMCS, University of Latvia. http://hdl.handle.net/20.500.12574/78.

Baklāne, A. (2023). Latviešu senākie romāni: aizmiršanas proporcija. *Punctum Magazine*, July 4. https://www.punctummagazine.lv/2023/07/04/latviesu-senakie-romani-aizmirsanas-proporcija/.

Bode, K. (2018). *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press. https://doi.org/10.3998/mpub.8784777.

Corpus analysis platform of the National Library of Latvia, 2021-2024. https://nosketch.lnb.lv/#open.

Frakturs, Crowdsourcing platform for teaching AI to read Latvian Fraktur script, 2019-2024. https://frakturs.lnb.lv.

Latvian Prose Counter, an interactive website for exploring the quantitative parameters of 19th and 20th-century Latvian prose fiction. https://proza.lnb.lv/en/.

Schöch, C., Patras, R., Erjavec, T., Santos, D. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, 1, 25. https://doi.org/10.3828/mlo.v0i0.364.

Znotiņš, A., Cīrule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies – The Baltic Perspective,* IOS Press, Vol. 307., pp. 183–189.

REST Services for Corpus management Annotation and SearcH

Alessandro Tommasi
University of Pisa, Italy
ale@ctrl-z-bg.org

Cesare Zavattari
University of Pisa, Italy
cesare.zavattari@gmail.com

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

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

Abstract

This paper presents a back-end software that offers a set of micro web services for the general-purpose management and search of text documents and annotations. Initially developed for a digital epigraphy project, the system focuses on integrating texts and lexicons represented in different paradigms. Nonetheless, the solution is designed to be general and adaptable across various domains.

1 Introduction

The need to digitally encode both primary and critical data in standardised or common formats is widely recognised across several cultural heritage fields, including epigraphy and historical linguistics.

In digital epigraphy, integrating various resources for use by both humans and machines is essential. Unfortunately, this integration remains underdeveloped even in many projects focused on studying ancient cultures through language. Most existing tools and projects, like the Epigraphic Database Heidelberg (EDH)[1] (Grieshaber, 2019) and iSicily[2] (Prag & Chartrand, 2019), primarily focus on the archaeological and historical aspects of inscriptions, but lack the interactivity required to link these resources with linguistic databases. Moreover, these initiatives often lack online, ready-to-use systems for creating, editing, or annotating the digitised materials, hampering collaboration and slowing online availability.

Solutions like EFES (Bodard & Yordanova, 2020) and Recogito (Barker et al., 2019) provide interesting useful tools but are text-centric and lack RESTful APIs, which we believe is crucial for ensuring the versatility of software in modern digital humanities projects.

In the context of ItAnt, an Italian collaborative research project aimed at integrating techniques and methodologies from the conventional study of epigraphic materials, computational lexicography, semantic web, and other digital humanities subfields[3], we aimed to complement the current landscape by providing a user-friendly web platform, DigItAnt, for creating and exploring LOD-compliant (historical) lexica, natively interlinked with digital editions of inscriptions, and other relevant language resources.

The DigItAnt platform aims to provide a technological solution that meets the needs of historical linguists, whose *modus operandi* includes investigating ancient cultures according to their linguistic documentation. A particular focus is placed on the creation of interlinked linguistic resources according to well-accepted representational models (such as XML TEI and RDF Ontolex-lemon). For details on the overall concept and on the data models adopted see our previous papers (Murano et al., 2023; Quochi, Bellandi, Khan, et al., 2022).

The platform consists of a system of independent software components implemented as a Service-Oriented Architecture. The back-end services are designed to be general, so as to be flexible and serve different use cases. The two main back-ends, the LexO-server and the CASH-server, manage lexicons

[1] https://edh.ub.uni-heidelberg.de/inschrift/suche(last accessed 2024/09/02)

[2] http://sicily.classics.ox.ac.uk/inscriptions/(last accessed 2024/09/02)

[3] *Languages and Cultures of Ancient Italy: Historical Linguistics and Digital Models* (ItAnt), https://www.prin-italia-antica.unifi.it/ (last accessed 2024/09/02)

and (annotated) textual documents, respectively. They both expose APIs based on the HTTP protocol and exchange data in JSON format. A set of other independent components than handle additional functionalities, such as Authentication and Authorisation Information (AAI), handled via an independent Keycloak server instance[4], which can be configured to allow for CLARIN Single Sign On (SSO). Further details about the platform are given in Quochi, Bellandi, Mallia, et al., 2022).

2 The Corpus management and Annotation Component

In this paper, we focus on CASH (**C**orpus, **A**nnotation, and **S**earc**H** server) whose primary responsibility is to serve as the back-end for managing text collections, annotations, and associated metadata. The system was developed to handle richly annotated document collections, including both primary texts and extensive metadata related to their historical and contextual information. Its native use case is to deal with a corpus of *EpiDoc* XML digital critical editions of archaic inscriptions. In addition to the annotated reconstruction of the inscribed texts, the corpus includes a set of contextual, historical, and descriptive metadata, following the practices of digital epigraphy (see Murano et al., 2023 for details on the corpus, and Fig. 4 in Appendix for a simplified sample of ItAnt Epidoc document).

CASH is designed to be modular and extensible in multiple ways, including document ingestion, annotation and metadata semantics, data export, and multi-level queries. The back-end services expose APIs documented via Swagger[5], and the source code is available as open source The source code, written in Java with a MySQL-based persistence layer, is available open source[6].

2.1 Document ingestion

CASH implements a customisable import module that allows users (i.e. project managers that install and set up the back-end systems) to specialise it for handling format-specific requirements. The module already supports three document formats (plain text, CoNLL-X, XML TEI EpiDoc), and is customisable to different models. It also ensures that while the "raw" document data is always available for retrieval, the information can be interpreted during import, creating metadata and annotations according to the specific use case requirements. Metadata can thus be included in the original document (e.g. in a heading statement at the beginning), as in the case for instance of XML TEI or CoNLL-U documents[7].

2.2 Corpus Management and Metadata

CASH organises documents into a file-system-like structure, allowing for easy metadata enrichment, annotation, and efficient searching. Specifically, the management of documents is exposed as a set of Create, Read, Update, and Delete (CRUD) operations, which is a standard practice for systems responsible for managing objects.

In CASH, metadata can be associated both with folders and individual documents, they can be typed and further enriched with additional features, and are available for querying and retrieval, see section 2.4 below. Metadata in fact are represented as a complex objects that have a main key/value structure plus any number of additional sub-features. Interestingly, features (i.e., both keys and values) are not predetermined, so that user clients can choose any to their willing, making the system general and flexible. Metadata values can be scalars, boolean, numerical, string, or composite types, i.e., lists of other key/value types. For example, EpiDoc files encode, among other types, information about the dimensions of the physical support that held the inscription represented in the document. During ingestion, the customised importing module creates a metadata entry that has as key "dimensions" whose value is itself a complex structure composed of a set of key/value pairs that represent all related information (represented by means of curly brackets in the example here below):

[4] https://www.keycloak.org/

[5] https://digitant.ilc.cnr.it/cash_demo/swagger-ui/index.html?configUrl=/cash_demo/v3/api-docs/swagger-config add /cash_itant/ before v3/

[6] https://github.com/DigItAnt/CASH-server

[7] While the use of the term "metadata" is arguably improper in computer science terms, in this case, it is still commonly used to convey that the data is not the text itself, but data about it.

```
``dimensions": {``precision": ``high", ``unit": ``cm", ``responsible":
DeBenedittis1980ID, ``width": "1.5",``height": "6.7"}
```

whose keys ("precision", "unit", "responsible", "width" and "height") are dictated by the original file encoding scheme. Further interpretation of the metadata and annotations, beyond its importing from the source document format, is left to the user client or to exporting modules. The core system makes no assumption as to metadata and annotation semantics, so as to maintain maximum flexibility. This way the back end may be used to serve different use cases. For instance, it could be used with a front-end specialised for revising or annotating (CoNLL) treebanks.

2.3 Annotation

Like metadata, annotations are represented as complex objects anchored to specific locations in a document's text via their character offsets. For example, part-of-speech tags or word-sense annotations can be attached to tokens or user-defined spans of text.

An annotation is represented with layers and values, which can be largely customised by the user (via the importer or the client). CASH does not impose a fixed set of annotation types, nor it limits the number of attribute-value features an annotation can be enriched with, thus allowing flexibility for different use cases. For instance, they can include Uniform Resource Identifiers (URIs), for facilitating linking operations to external companion resources. Annotations identify spans of unstructured text (potentially multiple spans). At minimum an annotation specifies the `layer`, which defines the type of information provided, and a `value`, i.e., the annotation content. Again, CASH ensures maximum flexibility and does not impose a predefined set of layers or values. For example, in a treebank-like scenario, it would allow a "part-of-speech" layer, with POS tags as possible values (e.g., *NN, VB, ADJ, ...*), whereas in a WSD scenario we might have a "Sense" layer with URIs or IDs identifying senses in an external resource like Wordnet as values. Also, more than one annotation is allowed for the same portion of text. By making the back-end indifferent to specific layers of annotations, as well as to other details, the specificity of the admissible layers and values can be totally determined by the controlling application. The notion of `token` is treated as special kind of annotation, i.e., it is the only layer explicitly known to (and therefore handled specifically by) the system, because it is the most common form of segmentation across many languages and many technological settings[8] Annotations (including tokens) may be further enriched with additional features, which again are entirely determined by the client on the basis of the specific use-case it serves. For all practical purposes, features attached to an annotation can be thought of as a $json$ object, and may be particularly useful in those cases where annotations refer to other annotations.

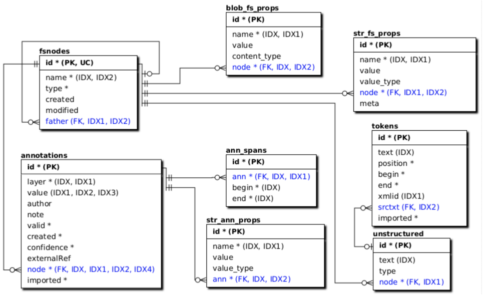


Figure 1: The schema for the underlying database.

Fig. 1 shows a fragment of the database that stores documents along with their associated tokens, metadata, and annotations. The schema includes a table representing "virtual" filesystem nodes (fsnodes), to which the document content, metadata properties (str_fs_props), and an unstructured text stream (the document's content) are linked. The unstructured content can be tokenised, with tokens anchored by

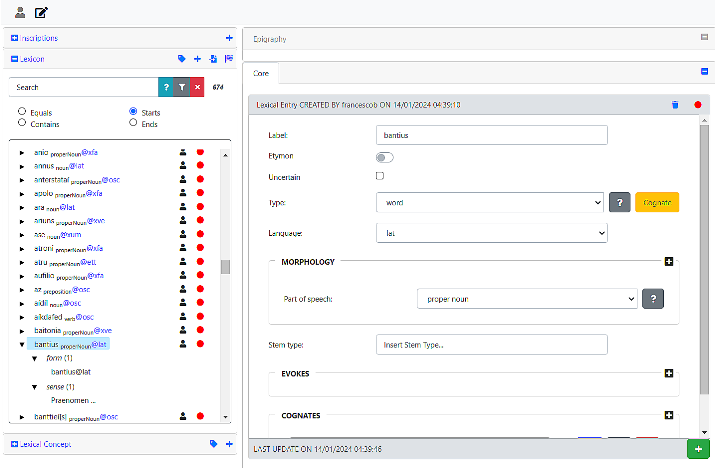
[8] We acknowledge the ongoing debate and criticism surrounding the concept and utility of tokens and tokenisation within the NLP community. However, this is beyond the scope of our discussion here and does not alter the fact that, as of now, tokenisation remains widely used.

their start and end character positions relative to the text. Annotations extend this concept, containing lists of spans (which can be non-consecutive) and associated metadata. Access to the APIs for creating, updating, and deleting documents, metadata, and annotations is secured with OAUTH-issued tokens, managed through a Keycloak installation. APIs for reading and querying the data are instead freely accessible. This way, data may be added and edited only by authorised users, whereas it can be visualised and searched openly.

2.4 Searching

CASH can use all the information persisted in its database for searching the documents. Searching is enabled along three axes: content, metadata, and annotations. User clients can implement functionalities to search for documents containing specific text sequences, metadata fields, or annotations such as tags, or IDs/URIs of standardised information encoded in external companion datasets. These three axes can be combined at will, enabling complex queries. To enhance usability, CASH employs the Corpus Query Language (CQL) (Jakubíček et al., 2010), but extends it to support multi-level searches. CQL has been chosen because it is indeed a well-known and widely used query language in corpus linguistics, likely familiar to many digital humanists, and it already possesses many features we required. Our extension specifically enhance CQL by enabling queries on metadata and supporting sub-token annotations, thus accommodating scenarios where CQL's typical assumption of word separation does not apply.

CASH is designed to be possibly deployed in association with different types of specialised front-end applications and user interfaces, serving different use cases. In practice, so far it is in use within the ItAnt project as one of the back-end serving two fron-end web applications: EpiLexO, for editing archaic lexicons and linking them to related inscriptions (in Fig. 2), and DigItAnt-search, for consuming and querying the DigItAnt data ecosystem (in Fig. 3). The only assumption that the software makes is that a text can be represented as a stream of UTF-8 characters. When this is the case, all other properties of the text (including formatting, typeface, syntactic, semantic layers and so forth) can be represented by annotations anchored to the text by spans over the UTF-8 stream.



○ Equals ● Starts
○ Contains ○ Ends

- anio $_{\text{properNoun}}$@xfa
- annus $_{\text{noun}}$@lat
- anterstataí $_{\text{properNoun}}$@osc
- apolo $_{\text{properNoun}}$@xfa
- ara $_{\text{noun}}$@lat
- ariuns $_{\text{properNoun}}$@xve
- ase $_{\text{noun}}$@xum
- atroni $_{\text{properNoun}}$@xfa
- atru $_{\text{properNoun}}$@ett
- aufilio $_{\text{properNoun}}$@xfa
- az $_{\text{preposition}}$@osc
- aídíl $_{\text{noun}}$@osc
- aíkdafed $_{\text{verb}}$@osc
- baitonia $_{\text{properNoun}}$@xve
- bantius $_{\text{properNoun}}$@lat
  - *form* (1)
    - bantius@lat
  - *sense* (1)
    - Praenomen ...
- banttieí[s] $_{\text{properNoun}}$@osc

Lexical Entry CREATED BY francescob ON 14/01/2024 04:39:10

Label: bantius
Etymon
Uncertain □
Type: word
Language: lat
Part of speech: proper noun
Stem type: Insert Stem Type...

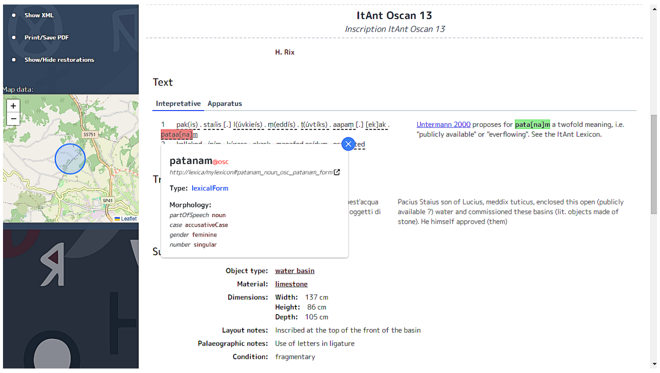
LAST UPDATE ON 14/01/2024 04:39:46

Figure 2: The editing web application

3 Relation with CLARIN-IT

CASH was developed as part of a CLARIN-IT supported project and is one of the key components of the DigItAnt platform[9], a CLARIN-IT/H2IOSC service. The software adheres to FAIR data principles, aligns with Open Science best practices, and is available as open source[10].

[9] https://digitant.ilc.cnr.it/epilexo_search_test/
[10] https://github.com/DigItAnt/CASH-server



ItAnt Oscan 13

Inscription ItAnt Oscan 13

H. Rix

Text

Intepretative Apparatus

1 pak(is) . staíis [.] l(úvkieís) . m(eddís) . t(úvtíks) . aapam [.] [ek]ak . pataa[na]m

Untermann 2000 proposes for pata[na]m a twofold meaning, i.e. "publicly available" or "everflowing". See the ItAnt Lexicon.

patanam@osc

http://lexica/mylexicon#patanam_noun_osc_patanam_form

Type: lexicalForm

Morphology:
partOfSpeech noun
case accusativeCase
gender feminine
number singular

Pacius Staius son of Lucius, meddix tuticus, enclosed this open (publicly available ?) water and commissioned these basins (lit. objects made of stone). He himself approved (them)

Object type: water basin
Material: limestone
Dimensions: Width: 137 cm
Height: 86 cm
Depth: 105 cm
Layout notes: Inscribed at the top of the front of the basin
Palaeographic notes: Use of letters in ligature
Condition: fragmentary

Figure 3: The exploration and search web application

4 Acknowledgments

This work was funded by the Italian Ministry of University and Research under the PRIN 2017 program and carried out within the project ”Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models” (PRIN 2017XJLE8J). It also received technological support from CLARIN-IT.

References

Barker, E., Isaksen, L., Kahn, R., Simon, R., & Vitale, V. (2019). Recogito. https://recogito.pelagios.org/help/about

Bodard, G., & Yordanova, P. (2020). Publication, Testing and Visualization with EFES: A tool for all stages of the EpiDoc XML editing process. *Studia Universitatis Babeș-Bolyai Digitalia*, *65*(1), 17–35. https://doi.org/10.24193/subbdigitalia.2020.1.02

Grieshaber, F. (2019). Epigraphic database heidelberg–data reuse options. Universitätsbibliothek Heidelberg. https://doi.org/10.11588/heidok.00026599

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast syntactic searching in very large corpora for many languages. *PACLIC*, 741–47.

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*(3), 1–14. https://doi.org/10.1145/3606703

Prag, J. R. W., & Chartrand, J. (2019). I. Sicily: Building a Digital Corpus of the Inscriptions of Ancient Sicily. In A. D. Santis & I. Rossi (Eds.), *Crossing Experiences in Digital Epigraphy: From Practice to Discipline* (pp. 240–252). De Gruyter Open Poland. https://doi.org/10.1515/9783110607208-020

Quochi, V., Bellandi, A., Khan, F., Mallia, M., Murano, F., Piccini, S., Rigobianco, L., Tommasi, A., & Zavattari, C. (2022). *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.

```
<?xml-model href="https://epidoc.stoa.org/schema/9.4/tei-epidoc.rng"
schematypens="http://relaxng.org/ns/structure/1.0"?>
[...]
<tei:teiHeader>
    <tei:fileDesc>
[...] <tei:editionStmt> <tei:edition> <tei:idno>ItAnt Oscan 2</tei:idno>
</tei:edition>
        <tei:editor> <tei:persName>Francesca Murano</tei:persName> </tei:editor>
      </tei:editionStmt>
      <tei:sourceDesc> <tei:msDesc> <tei:msIdentifier>
        <tei:settlement ref="https://sws.geonames.org/3180991">Campobasso
        </tei:settlement>
[...]
        <tei:institution ana="<url>">Soprintendenza Archeologia, Belle Arti e
        Paesaggio
            del Molise</tei:institution>
            <tei:idno>3974</tei:idno> </tei:altIdentifier>
            <tei:msName>Curse tablet from Monte Vairano</tei:msName>
            <tei:altIdentifier type="trismegistos">
                <tei:idno source="www.trismegistos.org/text/170774">TM 170774
                </tei:idno>
    [...]   </tei:altIdentifier> </tei:msIdentifier>
[...]
    <tei:physDesc>
     <tei:objectDesc> <tei:supportDesc> <tei:support>
        <tei:objectType ana="http://vocab.getty.edu/page/aat/300223016">tablet
        </tei:objectType>
[...]
        <tei:dimensions type="objectDimensions" unit="cm" precision="high"
           resp="#De_Benedittis1980a">
            <tei:height quantity="6.7">6,7</tei:height>
            <tei:width quantity="1.5">1,5</tei:width></tei:dimensions>
[...]   </tei:supportDesc>
        <tei:layoutDesc>
            <tei:layout columns="2" writtenLines="4">
            <tei:rs type="execution"
            ana="http://vocab.getty.edu/page/aat/300404794">
            engraved</tei:rs>
[...]   </tei:layout> </tei:layoutDesc>
     </tei:objectDesc>
     <tei:scriptDesc>
        <tei:scriptNote>
            <tei:rs type="writingSystem" subtype="alphabet" ref="#oscan-etruscan">
            Oscan National alphabet</tei:rs>
            <tei:rs type="wordDivision">punctuation</tei:rs>
            </tei:scriptNote> </tei:scriptDesc>
    </tei:physDesc>
    <tei:history> <tei:origin>
        <tei:origPlace type="composed">
          <tei:placeName type="ancient"
          ref="https://pleiades.stoa.org/places/438681">
          Aquilonia, Samnium</tei:placeName>
          <tei:placeName type="modern" ref="https://sws.geonames.org/3164966">
          Monte Vairano (Campobasso)</tei:placeName>
        </tei:origPlace>
 [...]
    </tei:history> </tei:msDesc> </tei:sourceDesc> </tei:fileDesc>
 [...]
    <tei:text>
     <tei:body>
        <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_2_1_1"/>
[...]
                <tei:name type="patronymic" xml:lang="osc-Ital-x-oscetr"
                xml:id="Osc_2_1_1_w_3" ref="#p1">
                <tei:expan><tei:abbr>tre</tei:abbr>
                <tei:ex>bieís</tei:ex></tei:expan></tei:name>
            </tei:ab>
[...] <tei:div type="translation" xml:lang="eng">
            <tei:p>Pacius Helvius son of Trebius | Statius Betitius [son of ...]
            </tei:p>
[...] </tei:div>
      <tei:div type="commentary" xml:lang="eng" resp="Francesca Murano">
 [...]</tei:div>
      <tei:div type="bibliography">
 [...]</tei:div>   </tei:body> </tei:text>
```

Figure 4: A sample of an ItAnt EpiDoc document. Most encoded parts are omitted for reasons of space. This serves simply as an exemple of the richness of information handled.

The Spanish INESData and TeresIA Projects as Potential Contributors to CLARIN

Elena Montiel-Ponsoda, Paula Diez-Ibarbia, Patricia Martín-Chozas
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain
emontiel@fi.upm.es, paula.diez@upm.es, pmchozas@fi.upm.es

Abstract

In this contribution, we provide an overview of two initiatives that have been recently launched in Spain for the creation and sharing of language resources and language technologies, amongst others, within the INESData and TeresIA projects. These two projects have received public funding from the NextGenerationEU recovery plan, through the Spanish Ministry for Digital Transformation and Civil Service (in the framework of its Recovery, Transformation and Resilience Plan). In both projects, language resources and language technologies play a crucial role, concentrating, mainly, on the languages spoken in Spain (Spanish, Catalan, Basque and Galician). The implementation of an integration mechanism with European infrastructures such as CLARIN is foreseen, to enable not only resource sharing but also resource discoverability and accessibility.

1 The INESData Project

INESData[1], the short form for "Infrastructure to Investigate Data Spaces in Distributed Environments at UPM" is a 2.5 years project which started in January 2023, funded by the Secretary of State for Telecommunications and Digital Infrastructures (SETELECO) at the Ministry for Digital Transformation and Civil Service, in the framework of the UNICO R&D Cloud program. The objectives of the INESData project are framed within the lines of work and fundamental principles proposed in the European Data Strategy The European Commission, 2020. This strategy envisions the creation of safe spaces for data exchange for both, public administrations and private companies, in federated infrastructures, and following the principles of sovereignty, privacy, transparency, security, and fair competition. According to this vision, data spaces will become the dynamic ecosystems to foster the creation of innovative data and services in Europe.

In this context, INESData aims to create an incubator of data spaces at a national level with the use of federated cloud and edge infrastructures, and governed by mechanisms that guarantee data sovereignty. The spaces created within the framework of the INESData project will be aligned with current data space initiatives in Europe, under the umbrella of the Data Spaces Business Alliance[2] (DSBA).

Specifically, the work to be carried out in the INESData project will be articulated around four main objectives:

1. To propose architectures aligned with European initiatives.
2. To develop basic components that support these architectures and facilitate the deployment of data spaces at national, regional or local levels, and in cloud-edge environments.
3. To develop and deploy value-added (micro)services, based on Artificial Intelligence (AI), paying special attention to the following: (a) natural language processing (NLP) services, (b) multimedia and multimodal data analysis services, and (3) services for the generation and exploitation of knowledge graphs in combination with (large) language models or LLMs.

[1] https://inesdata-project.eu/

[2] https://data-spaces-business-alliance.eu/

4. To deploy several data and service demonstrators, showcasing the applicability of the deployed infrastructure and developed technology.

With the ambition of highlighting the advantages of data spaces, we are currently working on the setting up of four vertical data spaces, that we plan to contribute to the corresponding sectoral European data spaces (also currently under development): language data space, mobility data space, media data space, and legal data space.

Concerning CLARIN and other European infrastructures, platforms and/or repositories of language resources, in the context of INESData we plan to contribute resources and services to those infrastructures, as well as enable the discoverability of and access to resources and services in some of those infrastructures. In the case of CLARIN, this will be achieved as part of and in coordination with the CLARIAH-ES[3] consortium and infrastructure.

In particular, the INESData language data space will rely on a catalogue of data and services, according to the recommendations detailed in the European Language Data Space[4] (ELDS) technical specifications, and analogous to catalogues of European NLP linguistic resources and services, such as CLARIN. The catalogue will be accessible both to people (with a user interface that allows faceted navigation, search by keywords, filters, etc.), and to machines (implementation of an API). In order to describe the resources included in the catalogue, the ELG-SHARE metadata schema (Labropoulou et al., 2020) will be taken as a starting point and adapted accordingly. Other W3C vocabularies or ontologies such as PROV to represent provenance information or DQV[5] (Data on the Web Best Practices: Data Quality Vocabulary) for data quality, will be employed as well.

Additionally, the catalogue will allow the registration of language resources and technologies by users of INESData to be deployed locally in the infrastructure or as RESTful web service, regardless of the server in which they are deployed.

Interesting for the CLARIN infrastructure will also be the legal and media data spaces. In the legal data space, INESData intends to create corpora in Spanish, and services for several purposes (for instance, classification of contract clauses). As for the media data space, multimedia a multimodal corpora are planned to be created and deployed in the data space.

2 The TeresIA Project

The name TeresIA (CSIC, 2023) stands for "terminologies in Spain and IA services". This project, which just started in January 2024, aims to provide an access point to linked terminologies in Spain, and IA services for term management. The project has received funding from the Secretary of State for Digitalization and Artificial Intelligence of the Ministry of Digital Transformation, in the framework of its Recovery, Transformation and Resilience Plan, and is framed in the Spanish Strategic Project (PERTE) of "The new economics of language" (España Digital, 2020).

The main objective of TeresIA is the development of digital tools based on AI, language technologies and data interoperability, to create a point of common and shared access to terminologies in Spain, and to generate, expand, reuse and apply terminological resources with great efficiency.

TeresIA will have two main services: one focused on access and recovery of terminological resources in a unified and optimal way through a metasearch engine; and another focused on extracting and validating terminologies using today's most advanced techniques, including AI and language models.

These services will allow users to execute different functions such as searching for the most appropriate terminology in an optimal way, linking the generated terminologies with other existing ones on the portal, or expanding existing terminologies. Other services will involve the semi-automatic extraction of terminologies from corpus and their subsequent validation. To develop these functionalities, the latest advances in language technologies (such as language models, AI, deep neural networks and transformers) and semantic web (such as linked data) will be used. Several state-of-the-art algorithms and methods relying on language models for term extraction (such as AttentionRank or MDERank) have been adapted

[3] https://www.clariah.es/

[4] https://language-data-space.ec.europa.eu/index$_e n$

[5] https://www.w3.org/TR/vocab-dqv/

for Spanish and configured in an easily executable manner. The initial results of their evaluation have been reported in (Calleja et al., 2024)

The infrastructure created will be used by the project partners in a series of use scenarios that will focus on the generation of terminologies in the legal field and/or enrichment of existing terminologies in the biomedical field. As a demonstration of the benefits of the terminologies created, these will be used in the management and recovery of content.

From a data governance point of view, the portal will act as a gateway to terminology in Spain, but respecting the licenses for the use of the resources available on it. At the moment of writing, several meetings have been organised with main terminology stakeholders in Spain, such as Termcat (the Terminology Centre for Catalan Language) or UZEI (a society attached to the Royal Academy of the Basque Language), amongst others. Additionally, such institutions are also directly involved in the TeresIA project through the Spanish Association for Terminology, AETER, one of the official partners in the project.

As for the portal functionalities, it will include user registration and management capabilities with different privilege levels, leaving the data query functionalities completely open and the addition, validation and management (validation/sanction) of resources limited to specific users. To facilitate the updating and maintenance of the tools, open source libraries will be used, as far as possible.

Ideally, TeresIA will also contribute data and resources to the CLARIN infrastructure to foster data sharing and reuse of the resulting terminology resources. As in the case of INESData, this will be achieved as part of the CLARIAH-ES consortium and infrastructure.

References

Calleja, P., Martín-Chozas, P., & Montiel-Ponsoda, E. (2024). Benchmark for automatic keyword extraction in spanish: Datasets and methods. *Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024*. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/293

CSIC. (2023). *El proyecto teresia recuperará y fomentará la terminología en español aplicando inteligencia artificial y conocimiento experto*. https://www.csic.es/sites/default/files/11diciembre2023_proyecto_TeresIA_101223%20(002).pdf

España Digital. (2020). Perte: Strategic projects for economic recovery and transformation. https://espanadigital.gob.es/en/measure/perte-strategic-projects-economic-recovery-and-transformation

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Rigault, M., Arranz, V., Choukri, K., Backfried, G., P'erez, J. M. G., & Silva, A. (2020). Making metadata fit for next generation language technology platforms: The metadata schema of the european language grid. *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:214713583

The European Commission. (2020). *Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions: A european strategy for data*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0066&from=EN

On the creation of multilingual information extraction workflows for literary-historical texts with generative LLMs

Tess Dejaeghere
Ghent University, Ghent
tess.dejaeghere@ugent.be

Pranaydeep Singh
Ghent University, Ghent
pranaydeep.singh@ugent.be

Els Lefever
Ghent University, Ghent
els.lefever@ugent.be

Julie Birkholz
Ghent University, Ghent
julie.birkholz@ugent.be

Abstract

Challenges regarding information extraction tasks from literary-historical texts arise from the unique nature of these texts, including historical languages, scarce benchmark datasets, variable annotation schemes, and digitization quality issues. Despite efforts to explore information extraction (IE) capabilities of open-source chat-based Large Language Models (LLMs), their application to literary-historical texts remains uncharted territory. The paper discusses the development of user-in-the-loop methodologies for applying and evaluating open-source chat-based generative models for named entity recognition, aspect-based sentiment analysis and relation extraction in literary-historical texts. Using a dataset of travelogues spanning English, French, Dutch, and German from the 18th to the 20th century as well as a corpus of letters from and to Guido Gezelle (19th century) as test cases, we present adaptable, multilingual workflows in Jupyter Notebooks. Our contributions include sharing datasets, notebooks and annotations for ABSA and NER via the CLARIN infrastructure and thus incentivizing engagement of the (digital) humanities community in assessing chat-based LLMs for information extraction from literary-historical texts.

1 Introduction

At the time of writing, discourse surrounding Large Language Models (LLMs) is reaching a fever pitch. Chat-based LLMs enable users to engage with models using natural language, revolutionizing communication paradigms and propagating a wide adoption of AI-tools across text-based tasks. Recent efforts have been dedicated to exploring and assessing generative LLMs' performance for information extraction (IE) tasks across various linguistic spaces and domains with variable results (Han et al., 2023; Li et al., 2023; Xie et al., 2023; Xu et al., 2023). At the time of writing however, the application and evaluation of LLMs as information extraction tools for literary-historical text material remains unexplored.

The reluctance of digital humanists to embrace generative LLMs as a new tool reflects their justifiably critical research attitude, given these models' propensity toward bias, the generation of unrequested or inaccurate content, irreproducible results, and privacy concerns. Literary texts are inherently extraordinarily challenging to annotate due to their subjective nature and stylistic properties, hindering standardization of both annotation and evaluation practices across the hermeneutics-driven (computational) literary domain (Bamman et al., 2019; Ehrmann et al., 2021; Ivanova et al., 2022; Kleymann & Stange, 2021; Rebora, 2023). From a technical point of view, treating works of literary-historical nature as data from which to extract information comes with specific methodological and technical challenges related but not limited to the presence of historical languages, the limited availability of domain-specific benchmark and training datasets, the wide variability of annotation schemes, and the variable quality of both digitization processes and off-the-shelf annotation tools (McGillivray et al., 2020; Moretti, 2013).

Fostering a working understanding of generative LLMs, their advantages, inner workings and potential pitfalls is an imminent need for the humanities community, given the rapidly increasing role of these models in education and research. To incentivize the implementation of such tools in practice and actively involve digital humanists in the red-hot debate regarding generative AI, we experimented with the application and evaluation of LLMs for IE-tasks such as NER, aspect-based sentiment analysis (ABSA) and relation extraction (REX) on 1) a literary-historical dataset of travelogues, and 2) a corpus of letters to and from the Flemish poet Guido Gezelle respectively. As a final step, our experiments were published as Jupyter Notebooks which can be adapted to other datasets in and beyond the literary-historical research domain.

2 Data

Our NER and ABSA workflows were developed using an annotated dataset of travelogues sourced from various online repositories. Eventually, we collected a dataset of 3,320 texts across the languages English, French, Dutch and German as shown in Table 1. A subset of 58 texts (approximately 5000 tokens/text) was manually labelled with entities (aspects) and their sentiment scores, and a second subset of 128 texts (approximately 500 tokens/text) was annotated with named entities. The named entities encompass concepts related to the environment, including person, location, organisation, fauna, flora, biome, human landforms, natural landforms, natural phenomena, weather and mythical creatures.

Language	**18thC**	**19thC**	**20thC**	**Total**
English	41	782	668	**1,491**
French	5	145	50	**200**
Dutch	25	92	242	**359**
German	972	218	80	**1,270**
Total	**1,043**	**1,163**	**897**	**3,320**

Table 1: Overview of languages contained in the travelogues corpus

Experiments for the relation extraction workflows were applied to a corpus subset of 50 letters in English, French and Flemish Dutch written to and by Guido Gezelle, a renowned Flemish poet and priest from the 19th century. These letters were gathered from the Gezelle Archive [1].

3 Methodology

Jupyter Notebooks are created for three separate tasks: NER, ABSA, and REX respectively.

As shown in Figure 1, multiple notebooks showcase the extraction of entities/aspects using 1) the HuggingFace API to launch calls to the open-source LLM mistralai/Mixtral-8x7B-Instruct-v0.1 in a zero-shot and few-shot prompting setting, 2) the Python bindings of the GPT4All-framework (Anand et al., 2023) to prompt and evaluate several open-source LLMs[2] locally in few- and zero-shot settings, 3) the NLP-package Flair (Akbik et al., 2019) through their off-the-shelf models as well as their zero- and few-shot model TARS, and 4) the NLP-package spaCy (Montani et al., 2023) through off-the-shelf models and the zero-shot model GliNER.

In order to evaluate our approach - prompts were incrementally made more complex and evaluated. starting out with a basic prompt with a task description, then adding a persona, an annotation guide, metadata (e.g.: the name of the author and the title of the book), and a set of few-shot examples. Evaluations are carried out by calculating F1 using our annotated dataset as a gold standard and the NLP-package nervaluate[3].

The output of the aspect extraction task can be transformed into a .csv-format and fed into the ABSA-notebooks. If sufficient annotated data is available, the user can also choose to use our machine learning-based training pipelines developed for the languages under consideration, which extract embeddings from language-specific models available on HuggingFace to serve as input for diverse machine learning classification architectures, including SVM, AdaBoost, Random Forest, and MLP classifiers.

For the relation extraction task - experiments were carried out using retrieval augmented generation (RAG) via two methodologies: in a first experiment, the user interface of the GPT4All application was

[1] https://gezelle.be/gezelles-brieven-een-participatieproject-1-algemene-pagina

[2] mistral-7b-instruct-v0.1, nous-hermes-llama2-13b and Meta-Llama-3-8B-Instruct

[3] https://pypi.org/project/nervaluate/

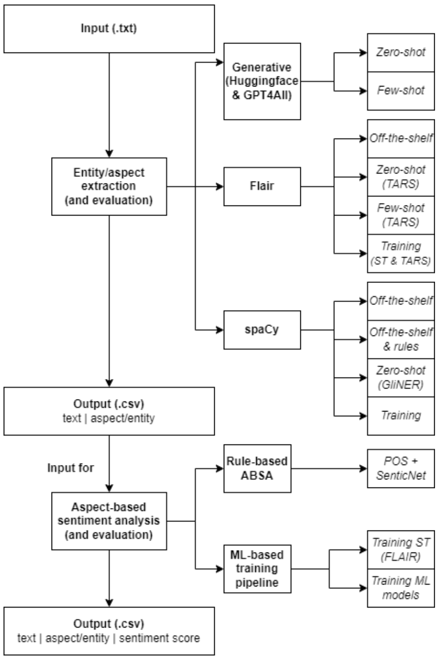


Figure 1: Overview of the Jupyter Notebooks created for the NER and ABSA tasks respectively.

used to create a vector database of background texts related to Guido Gezelle (a.o. biographies and Wikipedia pages) and a corpus of letters. Later, this vector database was coupled to an LLM and prompted for a list of relationships to the writer. In a second experiment, ChatGPT's API was used to create a RAG application. Code, data and annotations used for the experiments with ABSA and NER are already made open-source through our public GitHub page [4]. At the time of writing, our experiments with RAG are a work in progress and resources and results will eventually be published.

4 Expected contributions to the CLARIN infrastructure

The notebooks, datasets and annotations for the NER, ABSA and REX tasks serve as an example to incite further research in this domain, and will be made available through the CLARIN infrastructure and through the open science ecosystem via CLARIAH-VL (https://clariahvl.hypotheses.org/). They will also be shared as tools as part of the H2020 CLS INFRA – Computational Literary Studies Infrastructure Project (https://clsinfra.io/). The output can be freely adapted and used for training and research purposes in the Digital Humanities community and beyond. Our work represents a valuable step in the direction of user-in-the-loop workflows for lesser-resourced historical languages, and a pioneering effort in the exploration of open-source chat-based LLMs for NER and ABSA applied to the literary-historical domain.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In W. Ammar, A. Louis, & N. Mostafazadeh (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics (demonstrations)* (pp. 54–59). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-4010

[4] https://github.com/GhentCDH/CLSinfra

Anand, Y., Nussbaum, Z., Treat, A., Miller, A., Guo, R., Schmidt, B., Community, G., Duderstadt, B., & Mulyar, A. (2023). Gpt4all: An ecosystem of open source compressed language models. https://arxiv.org/abs/2311.04931

Bamman, D., Popat, S., & Shen, S. (2019, June). An annotated dataset of literary entities. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2138–2144). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1220

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2021). Named Entity Recognition and Classification on Historical Documents: A Survey [arXiv: 2109.11406]. *arXiv:2109.11406 [cs]*. Retrieved February 22, 2022, from http://arxiv.org/abs/2109.11406
Comment: 39 pages.

Han, R., Peng, T., Yang, C., Wang, B., Liu, L., & Wan, X. (2023, May). Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors [arXiv:2305.14450 [cs]]. Retrieved March 12, 2024, from http://arxiv.org/abs/2305.14450
Comment: 23 pages, version 1.0.

Ivanova, R., van Erp, M., & Kirrane, S. (2022, June). Comparing Annotated Datasets for Named Entity Recognition in English Literature. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3788–3797). European Language Resources Association. Retrieved February 5, 2024, from https://aclanthology.org/2022.lrec-1.404

Kleymann, R., & Stange, J.-E. (2021). Towards Hermeneutic Visualization in Digital Literary Studies. *DHQ: Digital Humanities Quarterly*, *2*(15). Retrieved December 7, 2021, from http://digitalhumanities.org:8081/dhq/vol/15/2/000547/000547.html

Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. (2023, April). Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness [arXiv:2304.11633 [cs]]. https://doi.org/10.48550/arXiv.2304.11633

McGillivray, B., Poibeau, T., & Fabo, P. R. (2020). Digital Humanities and Natural Language Processing: Je t'aime... Moi non plus. *Digital Humanities Quarterly*, *014*(2). https://www.digitalhumanities.org/dhq/vol/14/2/000454/000454.html

Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Landeghem, S. V., & Peters, H. (2023, October). Explosion/spaCy: V3.7.2: Fixes for APIs and requirements. https://doi.org/10.5281/ZENODO.1212303

Moretti, F. (2013). *Distant reading*. Verso.
Modern European literature: a geographical sketch – Conjectures on world literature – The slaughterhouse of literature – Planet Hollywood – More conjectures – Evolution, world-systems, Weltliteratur – The end of the beginning: a reply to Christopher Prendergast – The novel: history and theory – Style, Inc.: reflections on 7,000 titles (British novels, 1740-1850) – Network theory, plot analysis.

Rebora, S. (2023). Sentiment Analysis in Literary Studies. A Critical Survey. *Digital Humanities Quarterly*, *017*(2).

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., & Wang, H. (2023, October). Empirical Study of Zero-Shot NER with ChatGPT [arXiv:2310.10035 [cs]]. https://doi.org/10.48550/arXiv.2310.10035
Comment: Accepted to EMNLP 2023 (Main Conference).

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., & Chen, E. (2023, December). Large Language Models for Generative Information Extraction: A Survey [arXiv:2312.17617 [cs]]. https://doi.org/10.48550/arXiv.2312.17617

Can't See the Forest for the Trees: Tools and Services for Investigating Slovene Dependency Treebanks

Kaja Dobrovoljc
University of Ljubljana, Slovenia
Jozef Stefan Institute, Ljubljana, Slovenia
kaja.dobrovoljc@ff.uni-lj.si

Abstract

While the availability of high-quality syntactically parsed corpora has significantly increased, there remains a notable lack of infrastructural support for their effective linguistic exploration. This paper introduces a set of newly developed tools and services designed to address this gap in Slovene linguistics, developed in collaboration with the CLARIN.SI national consortium. Specifically, we present the Q-CAT tool for treebank annotation, the Drevesnik online service for treebank querying, the STARK tool for dependency tree extraction, and the CJVT Označevalnik interface for automatic treebank creation. All these tools are designed to work with dependency treebanks annotated following the Universal Dependencies scheme, ensuring a high degree of language independence and broad applicability.

1 Introduction

In addition to the well-established benefits to language technology (e.g. Zeman et al. (2018)), syntactically annotated corpora, i.e. treebanks, represent an equally invaluable data resource for research in linguistics and other language-based disciplines (e.g. Liu (2010)). Nevertheless, this methodological potential is yet to be fully exploited, which can partially be explained by the fact that investigating such complex type of data might present a challenge for researchers with little technical background. In this contribution, we present some recently developed tools and services that aim to overcome this infrastructural gap in Slovene linguistics.

2 Q-CAT Desktop Tool for Treebank Annotation and Analysis

The Q-CAT (Querying-Supported Corpus Annotation) tool (Brank, 2023) is a .NET Windows desktop application, which has been designed to support the manual annotation of reference corpora for Slovenian, such as SUK (Arhar Holdt et al., 2022), on various levels of linguistic description, including UD morphosyntax (de Marneffe et al., 2021). Users can utilize the tool to either browse existing annotations in an imported treebank (in TEI XML or CONLL-U format) or to create new annotations, as illustrated in Figure 1. Both modes benefit from the highly customizable settings interface, which enables users to define annotation layers and corresponding tagsets, associate them with specific colours for enhanced visualisation, and toggle them on or off as needed. Additionally, the tool has recently been enhanced to support the annotation of spoken data by linking transcriptions to audio recordings.

In addition to providing a platform for treebank annotation and browsing, the Q-CAT tool also features a search module that enables users to perform queries on top of these annotations, such as searching for multi-word expressions with specific syntactic structures. All matching sentences and tokens are displayed, counted and can be saved either as a subcorpus or a tab-separated list for further analysis and quantification. However, the queries are limited to a maximum of four tokens at a time, which represents a significant limitation for users interested in longer or more complex syntactic constructions.

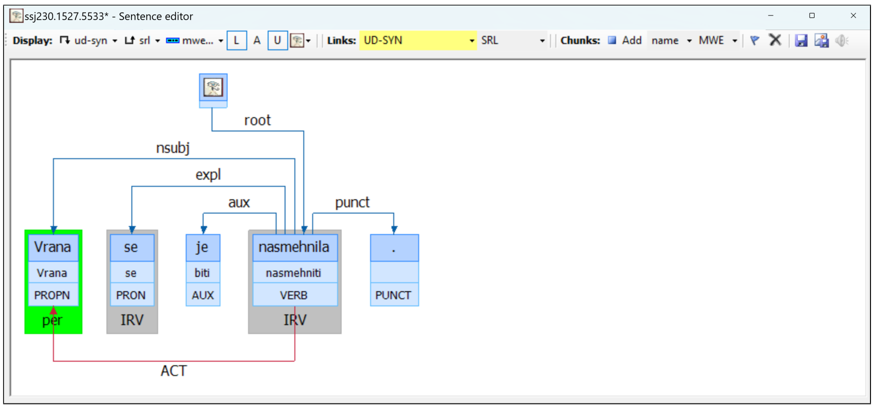


Figure 1: Example of a sentence in Q-CAT annotated for UD morphosyntax (e.g. *nsubj, PROPN)*, lemmas (e.g. *biti*), semantic roles (e.g. *ACT*), named entities (e.g. *per*) and multi-word expressions (e.g. *IRV*).

3 Drevesnik Online Service for Treebank Querying

To address the aforementioned limitation of Q-CAT querying interface, we recently developed Drevesnik (Štravs & Dobrovoljc, 2024),[1] an online service for querying Slovenian UD treebanks, which combines a powerful search engine on the one hand with user-friendly tree visualisations on the other. The tool is based on the open source Dep_search tool (Luotolahti et al., 2017), which was localized and upgraded so as to also support querying by MULTEXT-East morphosyntactic tags (i.e. the XPOS column in Slovenian CONLL-U data), randomize the order of the results and limit them to short sentences only (i.e. for demonstration purposes or teaching). In essence, the tool allows the users to describe the characteristics of the dependency structure of interest by means of a simple query language (explained and illustrated with Slovene examples), and returns the list of sentences matching the constraints in the form of dependency trees (graphs), as illustrated in Figure 2. The results are summarized by some basic statistics and can be downloaded either as a subcorpus (.conllu) or as a list of matched concordances.

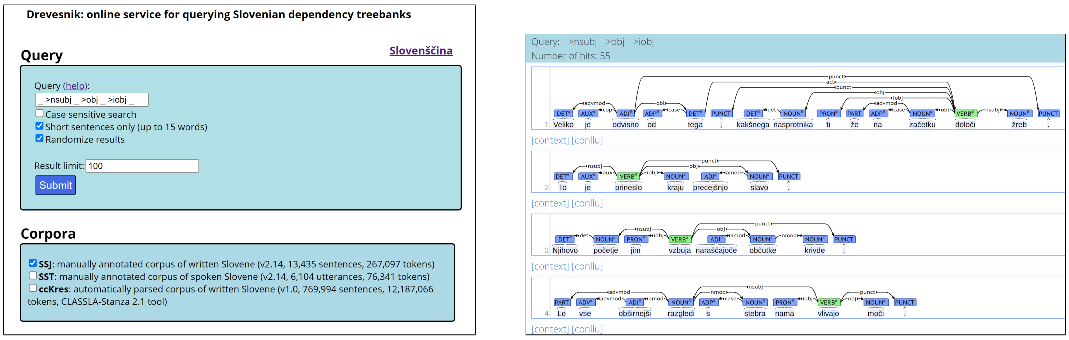
Three corpora are currently available: the SSJ UD treebank of standard written Slovene with 13,435 manually parsed sentences, the SST UD treebank of spoken Slovene with 6,104 manually parsed sentences, and the ccKres corpus with 769,994 automatically parsed sentences. However, any other treebank in CONLL-U can also be added in the future, as outlined in the project documentation on CLARIN.SI GitHub repository.[2]

4 STARK Tool for Dependency Tree Extraction

To complement traditional top-down, query-based approaches to treebank analysis with a more bottom-up, data-driven methodology, we developed STARK (Krsnik et al., 2024),[3] a Python-based command-line tool that generates frequency lists of dependency (sub-)trees from CONLL-U treebanks based on various customizible criteria. It allows for the extraction of both lexicalized and delexicalized trees (using either words or POS tags as tree nodes), which can then be differentiated (or not) with respect to surface node order, dependency structure or syntactic completeness. Users can also apply further restrictions based on the properties of the head node or specific label types.

The results from STARK are presented in a tabular file, listing the trees that meet the specified criteria in descending frequency order. This is illustrated in Figure 3, which shows the most frequent noun-headed trees in the English GUM UD treebank, i.e. the most common types of nominal phrases fea-

[1] https://orodja.cjvt.si/drevesnik/en
[2] https://github.com/clarinsi/drevesnik
[3] https://github.com/clarinsi/STARK



Drevesnik: online service for querying Slovenian dependency treebanks

Query

Slovenščina

Query (help):
_ >nsubj _ >obj _ >iobj _
□ Case sensitive search
☑ Short sentences only (up to 15 words)
☑ Randomize results

Result limit: 100

Submit

Corpora

☑ **SSJ**: manually annotated corpus of written Slovene (v2.14, 13,435 sentences, 267,097 tokens)
□ **SST**: manually annotated corpus of spoken Slovene (v2.14, 6,104 utterances, 76,341 tokens)
□ **ccKres**: automatically parsed corpus of written Slovene (v1.0, 769,994 sentences, 12,187,066 tokens, CLASSLA-Stanza 2.1 tool)

Query: _ >nsubj _ >obj _ >iobj _
Number of hits: 55

1 Veliko je odvisno od tega , kakšnega nasprotnika ti že na začetku določi žreb .

[context] [conllu]

2 To je prineslo kraju precejšnjo slavo .

[context] [conllu]

3 Njihovo početje jim vzbuja naraščajoče občutke krivde .

[context] [conllu]

4 Le vse obširnejši razgledi s stebra nama vlivajo moči .

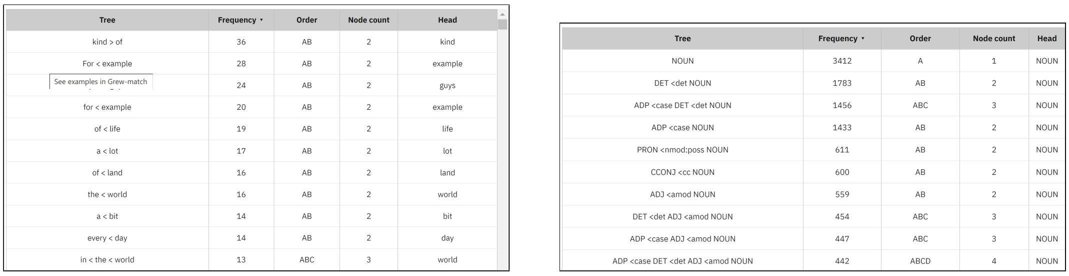
[context] [conllu]

punct
acl
punct
obj
iobj
advmod
cop
obl
case
det
advmod
case
obl
nsubj
DET AUX ADJ ADP DET PUNCT DET NOUN PRON PART ADP NOUN VERB NOUN PUNCT
nsubj
aux
punct
obj
iobj
amod
DET AUX VERB NOUN ADJ NOUN PUNCT
det
nsubj
iobj
punct
obj
amod
nmod
DET NOUN PRON VERB ADJ NOUN NOUN PUNCT
advmod
advmod
amod
nsubj
nmod
case
iobj
punct
obj
PART ADV ADJ NOUN ADP NOUN PRON VERB NOUN PUNCT

Figure 2: Example of a query (left) and matched trees (right) in Drevesnik.

tured in this corpus. In addition to the basic frequency counts, STARK computes various association measures, which are especially valuable for treebank-driven collocation extraction and lexical analysis. Additionally, it allows the users to compare their treebank to a second, reference treebank to identify treebank-specific syntactic or lexical phenomena by calculating various keyness scores.

Although the tool does not support additional tree visualization, it can print sentences containing the matched trees in the output. If the input treebank is an official UD treebank (recognized by the standard filename format), the output will include links to specific treebank examples in the Grew-match interface (Guillaume, 2021).[4] This feature is also available in STARK's demo online interface,[5] where clicking on a tree (i.e. a row in the output table) opens up a new tab with all relevant examples in the input treebank on Grew-match.



Tree	Frequency ▾	Order	Node count	Head
kind > of	36	AB	2	kind
For < example	28	AB	2	example
See examples in Grew-match	24	AB	2	guys
for < example	20	AB	2	example
of < life	19	AB	2	life
a < lot	17	AB	2	lot
of < land	16	AB	2	land
the < world	16	AB	2	world
a < bit	14	AB	2	bit
every < day	14	AB	2	day
in < the < world	13	ABC	3	world

Tree	Frequency ▾	Order	Node count	Head
NOUN	3412	A	1	NOUN
DET <det NOUN	1783	AB	2	NOUN
ADP <case DET <det NOUN	1456	ABC	3	NOUN
ADP <case NOUN	1433	AB	2	NOUN
PRON <nmod:poss NOUN	611	AB	2	NOUN
CCONJ <cc NOUN	600	AB	2	NOUN
ADJ <amod NOUN	559	AB	2	NOUN
DET <det ADJ <amod NOUN	454	ABC	3	NOUN
ADP <case ADJ <amod NOUN	447	ABC	3	NOUN
ADP <case DET <det ADJ <amod NOUN	442	ABCD	4	NOUN

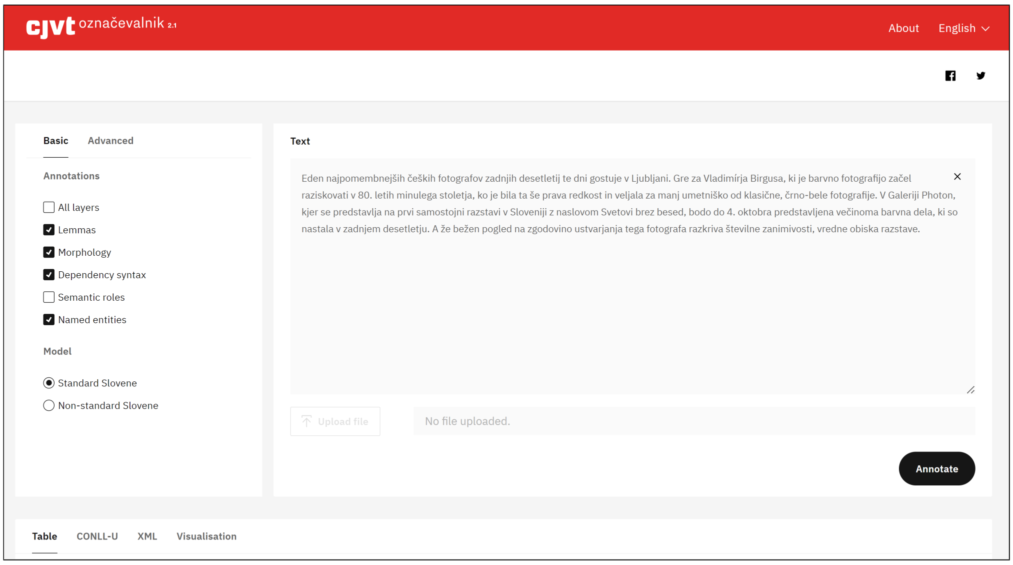
Figure 3: List of the most frequent lexicalized (left) and non-lexicalized (right) noun-headed trees in the English GUM treebank extracted with STARK.

5 CJVT Označevalnik Service for Automatic Text Annotation

CJVT Označevalnik is an online service for automatic grammatical annotation of Slovene text, designed not only to showcase the capabilities of such computational tools but also to assist non-technical researchers in creating new grammatically annotated corpora, such as those working on small-sized, genre-specific data. Inspired by the UDPipe web interface,[6] the tool assigns a variety of grammatical features to the user's input text, as illustrated in Figure 4, and produces (downloadable) results in four different formats: as tables, as CONLL-U and XML files, and as images visualised by the Q-CAT annotation tool (Section 2).

[4] https://universal.grew.fr/
[5] https://orodja.cjvt.si/stark/
[6] https://lindat.mff.cuni.cz/services/udpipe/

The service is based on the CLASSLA-Stanza NLP tool (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023),[7] which leverages a variety of language resources for Slovenian, such as the SUK training corpus, the Sloleks lexicon of inflected forms, CLARIN.SI word embeddings, and the Obeliks and ReLDI rule-based tokenizers. The CJVT Označevalnik service matches the latest version of the CLASSLA-Stanza tool, ensuring identical results, but offers a broader range of settings and output formats. The same interface design is also utilized in the CLASSLA Annotation Tool,[8] which, in addition to the models for processing (non-)standard Slovenian, also features models for other South Slavic languages.



Basic Advanced

Annotations

- ☐ All layers
- ☑ Lemmas
- ☑ Morphology
- ☑ Dependency syntax
- ☐ Semantic roles
- ☑ Named entities

Model

- ● Standard Slovene
- ○ Non-standard Slovene

Text

Eden najpomembnejših čeških fotografov zadnjih desetletij te dni gostuje v Ljubljani. Gre za Vladimírja Birgusa, ki je barvno fotografijo začel raziskovati v 80. letih minulega stoletja, ko je bila ta še prava redkost in veljala za manj umetniško od klasične, črno-bele fotografije. V Galeriji Photon, kjer se predstavlja na prvi samostojni razstavi v Sloveniji z naslovom Svetovi brez besed, bodo do 4. oktobra predstavljena večinoma barvna dela, ki so nastala v zadnjem desetletju. A že bežen pogled na zgodovino ustvarjanja tega fotografa razkriva številne zanimivosti, vredne obiska razstave.

No file uploaded.

Figure 4: Interface of CJVT Označevalnik service for automatic text annotation.

6 Conclusion

In this abstract, we introduced four recently developed or upgraded tools and services aimed at making dependency treebanks—complex data types that require advanced technical skills for effective analysis—more accessible to researchers in linguistics and digital humanities. For future work, we plan to develop additional training courses and materials to promote these tools within relevant communities, providing a broader range of compelling examples and use cases to encourage researchers to integrate these new methods alongside their traditional approaches.

Acknowledgments

This work was financially supported by CLARIN.SI and the Slovenian Research and Innovation Agency through the research project *Treebank-Driven Approach to the Study of Spoken Slovenian* (Z6-4617) and the research program *Language Resources and Technologies for Slovene* (P6-0411).

References

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., Pori, E., Terčon, L., Munda, T., Žitnik, S., Robida, N., Blagus, N., Može, S., Ledinek, N., Holz, N., Zupan, K., Kuzman, T., Kavčič, T., Škrjanec, I., . . . Zajc, A. (2022). Training corpus SUK 1.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1747

Brank, J. (2023). Q-CAT corpus annotation tool 1.5 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1844

[7] https://github.com/clarinsi/classla

[8] https://clarin.si/oznacevalnik/eng

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Guillaume, B. (2021, April). Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In D. Gkatzia & D. Seddah (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 168–175). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.21

Krsnik, L., Dobrovoljc, K., & Robnik-Šikonja, M. (2024). Dependency tree extraction tool STARK 3.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1958

Liu, H. (2010). Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, *120*(6), 1567–1578.

Ljubešić, N., & Dobrovoljc, K. (2019). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, 29–34. https://doi.org/10.18653/v1/W19-3704

Luotolahti, J., Kanerva, J., & Ginter, F. (2017, May). Dep_search: Efficient search tool for large dependency parsebanks. In J. Tiedemann & N. Tahmasebi (Eds.), *Proceedings of the 21st nordic conference on computational linguistics* (pp. 255–258). Association for Computational Linguistics. https://aclanthology.org/W17-0233

Štravs, M., & Dobrovoljc, K. (2024). Service for querying dependency treebanks drevesnik 1.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1923

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., & Petrov, S. (2018). CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–21. http://www.aclweb.org/anthology/K18-2001

Using Topics2Themes and Word Rain to visualise topics in Swedish news on climate change

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

Magnus Ahltorp
Language Council of Sweden,
Institute for Language
and Folklore
Stockholm, Sweden
magnus.ahltorp@isof.se

Abstract

The classic word cloud remains a popular visualisation technique, also to use for more advanced text exploration and comparison tasks. However, since the standard word cloud does not provide any support for these kinds of analytical tasks, we have created the Word Rain visualisation technique, which is a development of the classic word cloud. The Word Rain technique positions paradigmatically similar words close to each other on the x-axis, which makes it easier to identify semantic word clusters and to carry out comparison tasks. We have previously applied the technique on several different tasks, and we here show how the Word Rain visualisation can support a topical analysis of the text collection content. We first apply the topic modelling tool Topics2Themes to a collection of texts on the subject of climate change, and then use the Word Rain technique to visualise the automatically extracted topics. The Word Rain visualisation applied on the entire text collection provides an overview of its content, sorted according to paradigmatic similarity. When also creating focused word rain visualisations for the extracted topics, a visual semantic profile for each one of the topics is created, which supports the tasks of understanding and comparing topics. We have, thereby, here provided yet an example of how the Word Rain technique can be practically used for visualising and exploring texts.

1 Introduction

For visualising the content of a text collection, e.g. within digital humanities, the classic word cloud remains a popular technique. The word cloud does not provide any support for analysing text content, that is not provided by a simple word frequency list. However – despite this limitation – the word cloud is often applied for such tasks, e.g. to find relevant topics or word clusters in a text collection (Hicke et al., 2022). To address this problem, we have developed the Word Rain text visualisation technique (Skeppstedt et al., 2024), which provides a semantic sorting of the words. More specifically, instead of positioning the words in a random or alphabetical order, the words are ordered along the x-axis, with paradigmatically similar words being positioned close to each other. Thereby, the user is supported in the task of identifying clusters of semantically similar words, as well as in the task of comparing two graphs.

So far, we have mainly investigated the usefulness of the Word Rain technique for comparison of texts, e.g., to compare texts belonging to different text genres, or to study how the content of a text collection evolves over time. We have also applied the technique to provide an overview of small and topically focused text collections. For the topically focused texts, we have used a general language background corpus to be able to create a visualisation that emphasises words typical to the topics of the texts (Skeppstedt et al., 2024). However, we have not previously investigated the usefulness of the Word Rain technique for providing an overview of a text collection that instead is moderately diverse when it comes to the topics included. We hypothesise that topic modelling could be a useful method for visualising the content of such a text collection. In this study, we therefore do not only include a Word Rain visualisation applied on the entire corpus, but also Word Rain visualisations of the output of a topic modelling algorithm applied on the text collection. More specifically, instead of visualising the most prominent words in the entire corpus, we visualise the most prominent words in texts typical to the topics extracted.

The approach presented here is not the first study where the output of topic models is being visualised (Chaney & Blei, 2012; Sievert & Shirley, 2014), and the approach of using word clouds for visualising topic modelling output is also not unique (Liu et al., 2009). We are, however, not aware of any previous approaches where semantically motivated word clouds are used for this task.

2 Method

As the example for this study, we used the topic modelling output provided as a usage example by CDHUppsala[1]. The example had been constructed using a corpus of short Swedish news texts, from which climate change-related texts were extracted based on the keywords "climate change" and "climate crisis" (Skeppstedt, 2023). For extracting topics from the text collection, the Topics2Themes[2] tool (Skeppstedt et al., 2018) was used. In contrast to the visualisation of the corpus provided by CDHUppsala, we here did not take the temporal component of the text collection into account.

The topic modelling algorithm had generated 39 topics. For each topic, we extracted texts typical to the topic, and generated word rains for each topic based on these texts. We also generated a word rain graph for the entire corpus. As word prominence measure, we used tf-idf, together with a background corpus in the form of the Parole corpus, where each sentence was counted as a document for the tf-idf calculations[3]. We applied the same stop word list as had been used by the topic modelling algorithm (with the exception of the climate change keywords used for extracting the corpus) and we also added a few extra stop words. The word rains were configured to include the top 500 most frequent words, given that the word occurred at least 10 times in the topic-typical texts visualised, and to allow n-grams with a maximum number of five words. In all graphs, the five most typical words for each topic – according to the topic modelling algorithm – were underlined and marked in grey[4].

3 Results, discussion and conclusion

The Word Rain applied to the entire corpus (position **1A** in Figure 1) provides a semantically structured overview of the entire text collection, and makes it possible to identify clusters of words that are paradigmatically similar. For instance, starting from the left, it is possible to identify words in orange related to weather events, to climate change and its causes. A bit further to the right are places in nature (e.g. forest, the arctic) and species. When the colour starts turning into yellow, there are words related to water, draught and food. Then there is a green-yellow cluster of mostly geographical words. In the middle of the graph there is a cluster of mostly general words, but in the left part of this cluster there are many words describing change, and to the right (in blue) there are words related to children. The next word cluster in magenta starts with words related to research, then there are words related to organisations, politics and climate summits/agreements, and then countries. The graph then ends with a cluster of politicians/political parties, a cluster of names and finally a cluster mainly consisting of Swedish geographical names.

The word rain applied to the entire corpus thus supports the task of identifying clusters of paradigmatically similar words among those prominent in the text collection. This visualisation does, however, not provide any information regarding which of these prominent words often *co-occur* in the same news articles. For instance, from the first graph (**1A**), we can conclude that "parisavtalet" (the Paris Agreement) is a prominent word, but we need to consult the other graphs to find out with which other prominent words it often co-occurs. In graph **7A**, we can see that the word Paris Agreement co-occurs with words explaining climate change, in graph **2B** that it co-occurs with discussions on Amazonia, in **9C** with the IPCC-report, in **3D** with the topic of China, in **4D** in discussions on rich and poor countries, and finally in **6D** in relation to the U.S. and Donald Trump.

By displaying many topically focused word rains, we have transformed the task of identifying topics within a word rain to the task of comparing word rains. Such a transformation could also be done with a

[1] The example, which is based on a CLARIN node corpus (Språkbanken, Text: https://spraakbanken.gu.se/resurser/svt) is found here: https://github.com/CDHUppsala/topic-timelines.

[2] With the graphical user interface of Topics2Themes, texts extracted can be associated with manually identified 'Themes'.

[3] Språkbanken, Text: https://spraakbanken.gu.se/resurser/parole

[4] We used a pre-trained word2vec model for determining the paradigmatic word similarity: http://vectors.nlpl.eu/repository/

classic word cloud. However, while two classic word clouds are very difficult to compare, the semantic word ordering along the x-axis of the Word Rain makes the graphs optimised for comparison. We manually ordered the word rains generated, to emphasise the different visual profiles created by the semantic word ordering. The word rains in the left-hand side of Figure 1 all have a focus on the nature, with words – coloured in orange – that are positioned in the left part of the graphs. Example topics include temperatures (**2B**), snow (**3B**), floods (**4B**), wild fires and California (**5B**), heat waves and Europe (**6B**). Further down in the figure, we find topics on forests (**1B–2B**), water (**3B–4B**), plants and animals, e.g. invasive species, reindeer and tick-borne diseases (**6B-8B**). Continuing in the figure, we find the topics meat (**9B**), childbearing (**1C**), children (**2C**) and museums (**3C**), which all have a visual profile distinct from the topics on nature. We then have a number of visually similar topics about organisations and research: counties and municipals (**4C–5C**), research/climate-related research organisations (**6C–9C**), UN and EU (**1D–2D**). The figure then ends with topics on countries/politics/meetings/summits (**3D–8D**) – which all mainly have prominent words on the right-hand side of the graph – and a topic on climate activism and Greta Thunberg with a slightly different visual profile (**9D**).

We have here provided an example of how the Word Rain's semantic sorting can be used, not only for creating an overview of the entire text collection, but also for creating distinct visual profiles for the output of a topic modelling algorithm applied to the corpus. As a next step, we will continue to compare the approaches used here with other methods aimed at providing a corpus overview, both different types of visual overviews and purely text-based summaries.

Acknowledgements and tool availability

The code for generating word rains can be found at the GitHub page for Centre for Digital Humanities and Social Sciences, Uppsala (CDHU): https://github.com/CDHUppsala/word-rain. A web service for generating word rains is published by the Institute for Language and Folklore as a part of the CLARIN Knowledge Centre for the Languages of Sweden: https://wordrain.isof.se. Finally, the code for Topics2Themes can be found at the GitHub page for the Language Council of Sweden/the CLARIN Knowledge Centre for the Languages of Sweden: https://github.com/sprakradet.

The tools are developed within Swe-Clarin/The National Language Bank of Sweden (2017-00626), Huminfra (2021-00176) and InfraVis (2021-00181), all funded by the Swedish Research Council.

References

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing topic models. *Proceedings of the International AAAI Conference on Web and Social Media*.

Hicke, R. M. M., Goenka, M., & Alexander, E. (2022). Word clouds in the wild. *2022 IEEE 7th Workshop on Visualization for the Digital Humanities (VIS4DH)*, 43–48.

Liu, S., Zhou, M. X., Pan, S., Qian, W., Cai, W., & Lian, X. (2009). Interactive, topic-based visual text summarization and analysis. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 543–552.

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In J. Chuang, S. Green, M. Hearst, J. Heer, & P. Koehn (Eds.), *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). Association for Computational Linguistics.

Skeppstedt, M. (2023). Topics in Swedish News on Climate Change: A timeline 2016 – 2023. *CLARIN Annual Conference Proceedings*.

Skeppstedt, M., Ahltorp, M., Kucher, K., & Lindström, M. (2024). From word clouds to word rain: Revisiting the classic word cloud to visualize climate change texts. *Information Visualization*, *23*(3), 217–238.

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*.

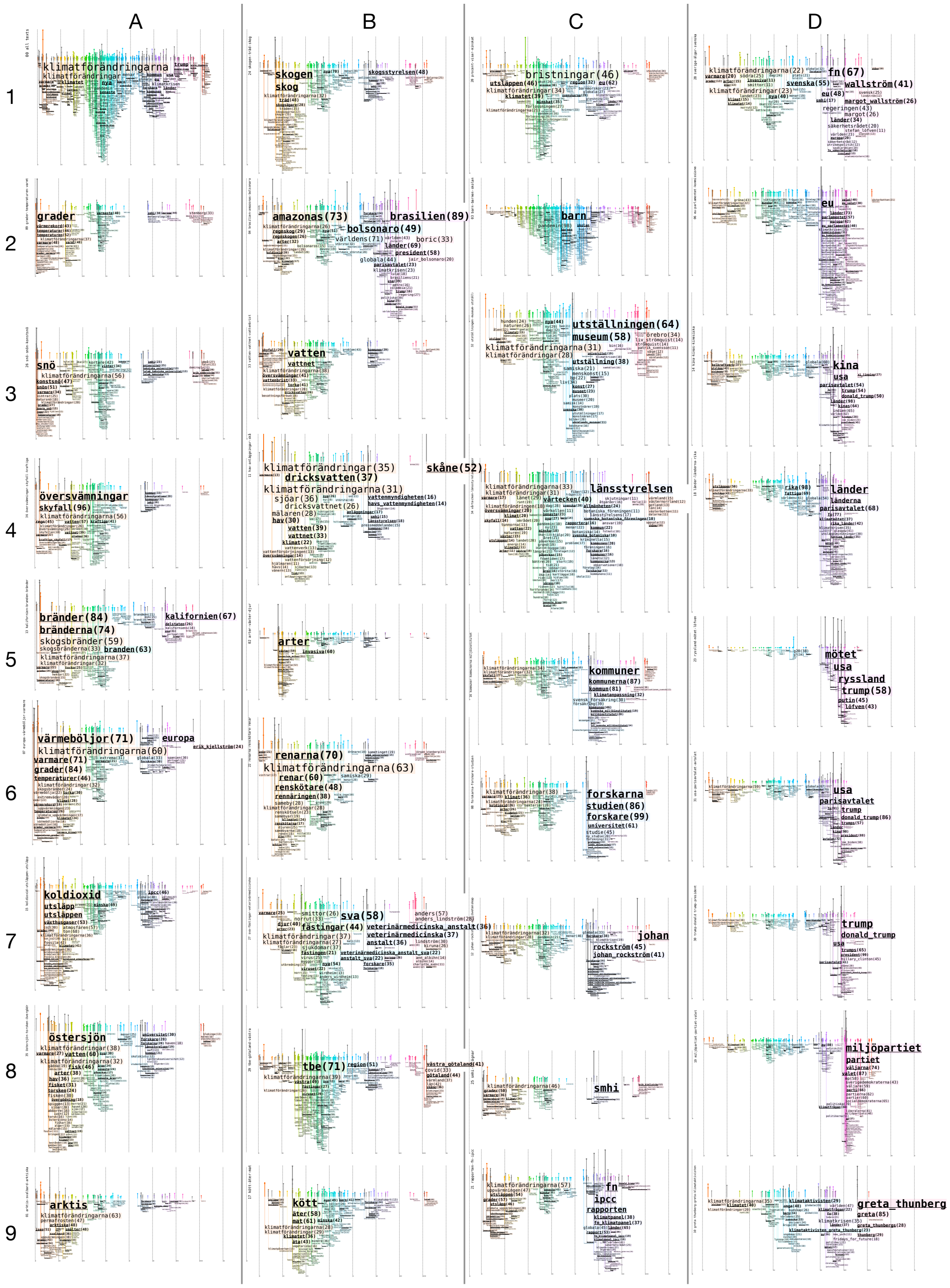


Figure 1: The word rain in the upper left corner is produced by visualising the 500 most prominent words in the entire corpus. The rest of the word rains visualise 35 of the automatically extracted topics. All graphs share the same semantic x-axis, making it possible to compare the semantic content of the different topics by comparing the x-axis positions where the most prominent words are located. (Letters and numbers are used in the text for referring to the graphs, e.g., the first graph is referred to as **1A**.)

Modeling events in Bulgarian: a Case Study

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

1 Introduction

The first version of Bulgarian event corpus (BEC) has been described in (Osenova et al., 2022). The corpus has been annotated on two levels: Named Entities and events with their roles. Here our focus are the events. On the level of events, the annotation scheme considers two models of abstraction - the conceptual resource FrameNet[1] and the ontology CIDOC-CRM.[2] These two sources were needed since the FrameNet for English is an already well-developed resource of frames while the ontology would give us the proper inheritance hierarchy and relations among the events. At the same time, the ontology does not cover all the events in the required granularity while FrameNet provides more specifics as well as various angles on them. However, the present frames of ours are utterly data-driven. This means that they were evoked by the data being annotated. Since the data within CLaDA-BG consists of mainly biographies and historical or ethnographic texts, the schema predominantly reflects the events of birth, death, occupation, charity, moving from one place to another, family relationships, some historical events, etc. The aim of the annotation of BEC is to support the extraction of facts to be included in the Bulgaria-centered Knowledge Graph. In this paper we discuss the annotation scheme in its part on the events with their roles, and propose some changes in it. We also report on modeling specific events from the point of active vs. passive perspectives. Non-surprisingly, the selected events for deeper inspection are **Birth**, **Death** and **Occupation**. We consider this work as a first attempt towards modeling events in Bulgarian from the variety of perspectives encoded in them. The resource has been developed within CLaDA-BG with the aim to support event extraction tasks for Bulgarian through training models and tuning them for respective domains and needs such as cultural heritage data, newsmedia data, criminal chronicles, etc. Our intention is to use the knowledge from the event corpus during the fine-tuning phase of LLM models. In this way we believe that better extraction of factual knowledge about Bulgarian history and culture as well as better indexing of data will be achieved.

2 The Annotation Scheme

There are various approaches towards event annotations. For example, Pan et al. (2006) use action vs. state division, aspectuality and reporting (quoted, unquoted), multiple events, negated events. In the same line, Vauth et al. (2021) use four types of events: changes of state, process, events, stative events and non-events. Tan et al. (2022) view the event corpus from the point of view of causality. Our approach is frame-based.

Following Minnema et al. (2022), we prepared a schema for additional annotation over the three selected events. Our scheme is simplified and we do not use any crowdsourced evaluation for detecting the focus in the sentence. Instead, we rely on the data itself and on the view of the annotators with the help of the overt linguistic triggers.

Our schema stipulates the following: a) divide the triggers into verbal and non-verbal types (No-Verbal), and consider only the verbal ones; b) look at the frame - if there is an agent role, then the

[1] https://framenet.icsi.berkeley.edu/

[2] https://www.cidoc-crm.org/

Event	Roles
Birth	**brought–into–life** (the newly born person) **parents** (the mother and father referred together) **mother** **father** **time** (the time of birth — usually it is a date, but can include hours, etc.) **place** (the birth place — usually the name of a city, country and/or hospital)
Death	**deceased** (the person who lost their life) **reason** (why the person died) **manner** (circumstances of the death) **related-event** (situation in which the death occurs) **place** (the place of death — usually the name of a city, country, etc) **time** (the time of death — usually it is a date, but can include hours, etc.)
Occupation	**agent** (person) **position** (what the job is of the agent) **area** (field of activity) **employer** (person or organization: in a broader sense, not necessarily formal) **related-event** (situation: event determining the occupation) **payment** (money or object: if the occupation requires such assets **goal** (situation: what the agent wants to achieve) **result** (situation or object: what was achieved) **place** (the place of work) **time** (the period of work)

Table 1: The roles of the three selected events as they are described in the annotation scheme.

event is active (*The killer shot the victim*) — ActiveInd; if not - it is passive (*He fell down the stairs*) — PassiveRo; c) if the frame has an agent role, then classify them either as an active voice (*The killer shot the victim*) or as a passive voice (*The victim was shot by the killer*) — ActivePas. No-Verbal cases happen when the trigger is a participle, Gerund or a noun. *Construct* is when there is not trigger, but event is recognized on the basis on the construct used in the text (*Ernest Miller Hemingway (July 21, 1899 – July 2, 1961)*). In this example the brackets indicate Hemingway's birth and death with the corresponding roles for the person that was born and was deceased. The times are explicitly represented, but there is no explicit trigger for these events.

Table 1 presents the three events that we discuss in the paper. The representation consists of a list of participant roles. Since the schema is data-driven, it reflects - although not exhaustively - the observations presented separately for each event. Thus, it can be seen that *Occupation* is an Agent-oriented event, *Birth* reflects only the coming into existence perspective and *Death* - only the victim perspective. While for *Birth* the closer examinations proved the predominance of this particular perspective, for *Death* the Agent one is also present and thus, the annotation scheme should be extended to cover it. The viewpoints depend on the source data being mainly biographies.

For the frequency of the three selected events within the corpus see Table 2. One step in modeling events that was called for but was not detected to be present in the first version of the resource is the stratification of the events. This means that they should follow a hierarchy since some events are more general while other remain more specific. Thus, events like *End* or *Beginning* are very general and can dominate a number of sub-events. For the End, these events are: *Leaving* defined as agents who are leaving a place or organization (CIDOC-CRM: E5 Event > E7 Activity > E86 Leaving; FrameNet: Process_End), *Dissolution* defined as an organization that ceases its existence (CIDOC-CRM:E5 Event > E7 Activity > E64 End of Existence > E68 Dissolution; FrameNet: Ceasing_to_be), *Death* defined as the death of a per-

Label	Occurences
ActiveInd	366
ActivePas	93
Construct	620
PassiveRo	77
No-Verbal	105

Table 2: Classification of the events Birth, Death, Occupation in categories: ActiveInd, ActivePas, Construct, PassiveRo, No-Verbal.

son (CIDOC-CRM: E5 Event > E7 Activity > E64 End of Existence > E69 Death; FrameNet: Death), *Destruction* defined as the demolition of some object without the creation of a new object (CIDOC-CRM: E5 Event > E7 Activity > E64 End of Existence > E6 Destruction; FrameNet: Destroying). For the *Beginning*, they are *Establishing* defined as the formation of groups, states, etc. (CIDOC-CRM: E5 Event > E7 Activity > E63 Beginning of Existence > E66 Formation; FrameNet: Cause_to_start), *Publication* defined as the published authorship of books, TV or radio programs (CIDOC-CRM: E5 Event > E7 Activity > E63 Beginning of Existence > E12 Production; FrameNet: Publishing), *Creation* defined as the invention of a myth, an idea, a rule, a book, a movie (CIDOC-CRM: E5 Event > E7 Activity > E63 Beginning of Existence > E65 Creation; FrameNet: Creating). This stratification step would also provide insights into the specific perspective. For example, in the *Birth* and *Death* events it should be distinguished between being born and giving birth as well as being dead, dying or causing death. Another thing is the need of event-triggers classification per event. Below the observations for the three selected events are presented.

3 Observations over Selected Events

Birth. This event shows the shortest list of triggers and respectively, of trigger types. It has registered only the type *Has-Agent (PassiveRo)*.

This event shows the specificity of the event corpus with respect to its perspective. In the data the act of birth is present as ‘being born’ but not as ‘giving birth’. This means that the perspective is shown for the one coming into existence and unfolding their life events but not for the one giving birth and taking care of the family. However, if the corpus is changed, this dominant perspective might change as well.

Linguistically, three forms were used in Bulgarian: the reflexive passive (*se razhda*), the participle passive (*e roden*) and an already archaic variant of the participle passive (*e rodom*). The first two are translated as ‘to be born’. The third one is rather translated as ‘to be born somewhere’.

Death. For the *Has-Agent (ActiveInd or ActivePas)* situation the following triggers have been registered in Active voice: *ubivam* (to kill), *zastrelvam* (to shoot), *samoubivam se* (to suicide), *izbivam* (to slaughter), *besya* (to hang someone), *likvidiram* (to liquidate), *nabivam na kol* (to impale), *slagam kray na zhivota si* (to end one’s life). This perspective provides general triggers like *ubivam* (to kill) or *likvidiram* (to liquidate) as well as more specific ones that show the manner of killing — *zastrelvam* (to shoot), *besya* (to hang someone), *nabivam na kol* (to impale). An always interesting case is the perspective of *samoubivam se* (to commit suicide) and *slagam kray na zhivota si* (to end one’s life) where the killer and the victim are the same person. Linguistically, it can be seen that not only verbs are used, but also multiword expressions of various kinds: *nabivam na kol* (to impale), *slagam kray na zhivota si* (to end one’s life). Thus, some verbs or verbal MWEs add to the event also the manner of death. For the *Has-Agent (PassiveRo)* situation the following triggers have been registered in Passive voice: *sam ubit* (be murdered), *sam obezglaven* (be beheaded), *sam zastrelyan* (be shot), *sam osaden na smart* (to be sentenced to death), *sam obesen* (to be hung), *sam likvidiran* (to be liquidated), *sam ekzekutiran* (to be executed), *padam ubit* (to fall shot), *sam izbit* (to be slaughtered), *sam udushen* (to be strangled), *sam umartven* (to be put to death), *sam izgoren zhiv* (to be burnt alive). In this cases the focus is on the result of the event and it is more important than the detailed description of it. Here predominantly the participle passive has been used. In general, the passive forms are forms of the active verbs from the previous group. For the *Lack-Agent* (ActivePas) situation the following triggers were registered: *umiram* (to die), *zagivam* (to die), *sam umryal* (to be dead), *zagubvam zhivota si* (to lose one's life), *pochina* (to pass away). Here the triggers refer to the general concept of dying where only *zagivam* (to die) can imply death caused by an accident or in a battle/at war. The *Death* event registers all the perspectives and is rich in triggers. It can be seen that most often the passive perspective on the *Has-Agent* situation is provided. This means that the perspective of the victim prevails over the perspective of the killer/murderer. The viewpoint depends on whose biography this is - the victim's one or the murderer's one. This again reflects the specificity of the corpus where mainly the life of famous historical persons who sacrificed their lives for their country is described.

Occupation. For the *Has-Agent (PassiveRo)* situation the following triggers have been registered in Active voice: *general*: *rabotya (kato)* (to work as), *zanimavam se s* (to deal with / to work), *sluzha* (to serve), *otgovaryam (za)* (to be responsible for), spetsializiram (to specialize), *pravya kariera* (to do career), *imam za zadacha* (to have a task); *specific-verb*: *rakovodya* (to manage), *oglavyavam* (to head), *koordiniram* (to coordinate), *prepodavam* (to teach); *specific-copula-noun*: *sam advokat* (to be a lawyer), *sam targovets* (to be a merchant), *sam glaven redaktor* (to be a chief editor).

For the *Has-Agent (PassiveRo)* situation the following triggers have been registered: *rakovodi se* (to be managed), is chaired (to be chaired), *prodava se* (is sold), *oglavyava se* (is headed). For the *Lack-Agent* situation there are no registered usages. This event is wider in its coverage. We are in the process of specializing it such that for example the 'position appointment' will be classified as another semantic group. However, at the moment this event is very rich in both - triggers and semantics. Also, the active Agent frames predominate. The triggers in these frames are verbs or multiword expressions. They can have a more general reference like *rabotya (kato)* (to work as), *zanimavam se s* (to deal with / to work), *sluzha* (to serve), or more specific one. The latter can be expressed in two ways - by a verb or a multiword expression, or by a copula-noun predicate. Typically the verbal and copula-noun expressions are paraphrases of the same content - *redaktiram* (to edit) vs. *sam glaven redaktor* (to be a chief editor). The copula-noun predicate seems to be used more often that the verbal predicates.

4 Conclusions

At the moment our work is focused on the following activities: a) re-classification of the events and thus, event hierarchy enrichment; b) improving the existing co-reference chains and c) checking the scope and labels of the participating named entities.

In this paper we report our first steps towards the definition of various perspectives over the events in BEC. The inclusion of differing perspectives would facilitate tasks like sentiment and stance analysis and would contribute to the critical thinking.

Our future work would go to at least three directions. The first is the gathering of texts with more elaborate sub-event types (from instructions, recipes, routines, etc.). The second is the gathering of crowdsourced event perspectives. The third is training of LLMs where the event knowledge is incorporated within the fine-tuning phase.

Acknowledgements

The reported work has been partially supported by CLaDA-BG, *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH*, funded by the Ministry of Education and Science of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure).

References

Minnema, G., Gemelli, S., Zanchi, C., Caselli, T., & Nissim, M. (2022, May). SocioFillmore: A tool for discovering perspectives. In V. Basile, Z. Kozareva, & S. Stajner (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics: System demonstrations* (pp. 240–250). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-demo.24

Osenova, P., Simov, K., Marinova, I., & Berbatova, M. (2022, June). The Bulgarian event corpus: Overview and initial NER experiments. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 3491–3499). European Language Resources Association. https://aclanthology.org/2022.lrec-1.374

Pan, F., Mulkar, R., & Hobbs, J. R. (2006, May). An annotated corpus of typical durations of events. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the fifth international conference on language resources and evaluation (LREC'06)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2006/pdf/234_pdf.pdf

Tan, F. A., Hürriyetoğlu, A., Caselli, T., Oostdijk, N., Nomoto, T., Hettiarachchi, H., Ameer, I., Uca, O., Liza, F. F., & Hu, T. (2022). The causal news corpus: Annotating causal relations in event sentences from news.

Vauth, M., Hatzel, H. O., Gius, E., & Biemann, C. (2021). Automated event annotation in literary texts. In M. Ehrmann, F. Karsdorp, M. Wevers, T. L. Andrews, M. Burghardt, M. Kestemont, E. Manjavacas, M. Piotrowski, & J. van Zundert (Eds.), *Proceedings of the conference on computational humanities research, chr2021, amsterdam, the netherlands, november 17-19, 2021* (pp. 333–345, Vol. 2989). CEUR-WS.org. http://ceur-ws.org/Vol-2989/short_paper18.pdf

Preserving Privacy in Small Communities: Tailored Anonymization Techniques for Icelandic Conversational Data

Elena Callegari
University of Iceland, Reykjavík
ecallegari@hi.is

Agnes Sólmundsdóttir
University of Iceland, Reykjavík
ags46@hi.is

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

Abstract

We examine the challenges and methodologies of anonymizing a dataset of Icelandic conversations, emphasizing the need for language-specific strategies due to Iceland's small, interconnected population and the morphological richness of the language. We discuss the importance of preserving grammatical elements such as case and gender to maintain data utility for linguistic research. The study proposes an anonymization technique that balances data utility with privacy, resorting to pseudonyms that match the original phrase's linguistic properties to protect individual identities while preserving the structural integrity of the Icelandic language.

1 Introduction

In the era of data-driven decision-making, the collection and analysis of conversational data have become essential for advancing linguistic research and possibly improving language technologies. However, the use of such data raises significant privacy concerns, particularly in the context of small, tightly-knit communities where individuals are easily identifiable even from seemingly innocuous information. This paper explores the anonymization of a dataset of audio-recorded Icelandic conversations, underlining the necessity of language-specific anonymization protocols to protect participant privacy while maintaining the utility of the data.

The conversations in question have been audio-recorded with the intention of creating the very first Icelandic Dementia corpus. We have been collecting speech samples from Icelandic individuals suffering from various stages of Alzheimer's Disease (AD) as well as from healthy, age-matched individuals Callegari et al., 2023, 2024. We plan to release the transcriptions of the conversations 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.

Anonymization involves processing personal data to remove or obscure identifying details, ensuring that individuals cannot be identified directly or indirectly by the retained data. Anonymization helps mitigate the risks of unauthorized data re-identification which could lead to privacy invasions, discrimination, or other forms of harm. Effective anonymization allows researchers to share and analyze datasets without compromising individual privacy, thereby adhering to ethical standards and legal requirements, such as Europe's General Data Protection Regulation (GDPR). Iceland, with its population of roughly 370,000, exemplifies a scenario where simple anonymization methods, such as merely removing proper names from data files, may not suffice. The Icelandic language is used by a relatively small speaker community. In such environments, these elementary anonymization approaches might leave sufficient linguistic and contextual clues that could potentially lead to the identification of individuals. This risk is particularly high in cases where unique personal references, local dialects, or specific sociolects are prevalent.

Moreover, the morphological richness of Icelandic—characterized by a complex system of inflections and derivations—means that anonymizing content without distorting linguistic structures requires careful consideration. Preserving the grammatical integrity of the language is essential for ensuring that the anonymized data remains valuable for linguistic research and the development of natural language processing tools.

2 Data to Anonymize

In the broader scope of data privacy and protection, certain types of information commonly require anonymization across various contexts and geographies. These include personal identifiers, such as full names, addresses, social security numbers; these are always anonymized to prevent the straightforward recognition of an individual's identity. Information that is also generally anonymized are health records, financial data, and personal communications, i.e. any sensitive information that could impact an individual's privacy and security if exposed.

In smaller countries or communities, the effectiveness of these basic anonymization techniques may diminish due to the increased likelihood of identifying individuals through indirect means. For example, in small communities, certain cultural practices, local events, or affiliations (e.g., membership in specific local organizations) can serve as identifiers. In Icelandic data, references to participation in local festivals, or membership in less-common local clubs could inadvertently reveal someone's identity. Moreover, while anonymizing a city's name might not be necessary for countries with larger populations, in Iceland, the name of a town or district might need to be anonymized due to the small number of inhabitants and the resulting ease of individual identification. Iceland also has a notable interest in genealogy, facilitated by extensive public and private records that trace family genealogies. Any data hinting at familial connections or lineage, such as particular names or patronymics, which might be benign in larger populations, could lead to individual identification in Iceland. In smaller or more specialized professional communities, detailing someone's educational background or employment history (specific roles, small or niche industries) can be particularly revealing. In Iceland, mentioning a person's role in a specific sector, like fisheries or geothermal energy, might narrow down the identity of individuals far more than in larger economies. Moreover, in small populations, even anonymized health data might be re-identifiable if it includes rare health conditions or treatments that are unique to a few individuals within the community. Finally, in datasets with fine-grained demographic segmentation, details that might be individually harmless, such as gender, employment, specific names of towns and festivals, can be used in combination with other data to re-identify individuals. Consider for instance examples (1) and (2):

(1)
Ég var alin upp á Hólmavík hjá foreldrum mínum, Árna og Jónu,
I was raised.FEM in Hólmavík at parents.DAT mine.DAT, Árni.DAT and Jóna.DAT

'I was raised in Hólmavík at my parents, Árni and Jóna.'

(2)
og pabbi minn var læknirinn í bænum og mamma vann í móttökunni hjá honum.
and dad.NOM mine.NOM was doctor-the.NOM in town-the.DAT and mom.NOM worked in recept

'and my dad was the town doctor and my mom worked at his reception.'

In datasets of languages spoken by larger communities, the personal details provided in examples (1) and (2) might seem innocuous, yet in a society as small as the Icelandic one, such a combination of geographic, familial, and occupational information may be enough to lead to individual re-identification. This is especially the case for our Dementia dataset, given that we interviewed a very specific demographic group (individuals aged 60 to 80 at the time of the interview, some of whom had a neurocognitive condition). Moreover, part of our interview protocol consists in asking participants to recall their childhood home; this often lead to descriptions of family members, and to mentions to schools, specific places and organizations.

3 Icelandic Morphology & Anonymization Strategies

One of the defining characteristics of the Icelandic language is its inflectional morphology. Icelandic has four cases: nominative, accusative, dative, and genitive. Nouns in Icelandic agree in gender (masculine, feminine, neuter) and number (singular, plural). Icelandic verbs are conjugated according to mood, tense, voice, person, number, and gender. The richness of this system is a significant aspect of the language's morphology. The morphological complexity of Icelandic has direct implications for data anonymization, particularly when considering the need to maintain linguistic integrity for research purposes. Case usage in Icelandic can reveal subtle demographic or sociolinguistic patterns. For instance, Callegari et al., 2024 have shown variations in the use of the dative case across different age groups of Icelandic speakers. Ideally, anonymization strategies should therefore preserve case information to allow for the study of such linguistic phenomena without compromising the privacy of the individuals involved.

Recently, numerous studies focusing on text anonymization across various languages have emerged (e.g. Francopoulo and Schaub, 2020; Mozes and Kleinberg, 2021; Strathern et al., 2020; Adams et al., 2019). Currently, there is no default anonymization method, as the choice of method varies across different fields of research and is dependent on the intended purposes of the data. Most studies suggest the following four key requirements for anonymizing text before publication: (i) ensuring the anonymity of participants and individuals mentioned in the text, (ii) allowing in-house semantic data analysis and language analysis through NLP, (iii) proof that an anonymization has taken place, and (iv) the method should be applicable for different European languages (Francopoulo and Schaub, 2020). Several options to anonymize conversational transcripts exist; what can sometimes be challenging is ensuring that all of the four requirements listed above are fulfilled at the same time.

One particularly straightforward method for anonymizing conversational data is to fully redact sensitive information by completely removing personally-identifying information, and replacing it with an "X" token. This can be suitable for the public release of government documents without secondary analysis. However, this method does not fulfill the aforementioned requirement number (ii), as important linguistic information such as semantic cohesiveness, syntax and other lexical properties needed for in-depth linguistic analysis would be lost (Mozes and Kleinberg, 2021).

Another method mentioned in Strathern et al., 2020 is aggregation, where identifiable information units are coarsened or aggregated by creating classes or categories, e.g. replacing someone's age with age classes, or replacing a specific person's name with a relevant but vague role, such as "student" or "employee". Aggregation therefore keeps some semantic cohesiveness and can be used for secondary analysis, i.e. sentiment analysis and topic modelling, but to a limited extent. The drawbacks are that, as in the first anonymization method, too much linguistic information is lost for NLP purposes.

A third and frequently mentioned method is the use of pseudonyms, which refers to renaming identifiable units, such as people, institutions etc. This can be done in two ways: either by using anonymous place-holders (e.g., 'Person-Name', 'City-Name' etc.) or by using unique identifiers, such as choosing a random name with comparable properties. For example, one could replace the name *Björk Gumundsdóttir* with a name like *Ösp Davísdóttir*: both are female names, and both feature a similar number of syllables and hence have a comparable length. The latter method, recommended by Aldridge et al. (2010), is considered more suitable for linguistic data analysis as the chosen pseudonyms match all linguistic properties of the original utterance.

3.1 Current Practices in Icelandic Data Anonymization

In our collaboration with other researchers working on anonymization for Icelandic corpora, several practical strategies have been highlighted. For example, in the court document corpus published by Clarin IS (Barkarson et al., 2022), a Named Entity Recognizer (NER), MIM-GOLD-NER, was used to identify and replace personal names with strings encoding the first letter, gender, and case marking. While this approach preserved grammatical information relevant to Icelandic, the accuracy of the model in recognizing foreign names and professions remained a challenge. Additionally, place names, streets, and organizations were anonymized using similar approaches, although the NER used in this project was not specifically designed to handle foreign names effectively.

These experiences suggest that for certain corpora, particularly those that involve sensitive legal or clinical information, a combination of Named Entity Recognition and Part-of-Speech (PoS) tagging could be an effective strategy for ensuring linguistic integrity.

4 Our Chosen Anonymization Practice

The corpus we will release will contain manually transcribed speech samples elicited from individuals of Icelandic nationality who are aged 60 to 80 and who are either healthy or suffering from various degrees of Alzheimer's Disease. Seeing as we are working with sensitive information, e.g. clinical diagnostic information, anonymity is of the utmost importance. However, as the main focus of our study is on the specific effects AD can have on language production, it is important not to lose any linguistic information relevant to the study.

The initial step in the anonymization of our data will involve establishing a comprehensive list of basic entity types that could contain identifiable information. This includes full names, addresses, social security numbers, and other direct identifiers. Additionally, as outlined in section 2, we will identify a subset of entity types that may be traceable within smaller communities like Iceland. This subset will potentially include names of organizations, schools, cities, towns, regions, and employment roles, among others. Following the creation of this detailed inventory of elements requiring redaction, we will proceed to anonymize our dataset.

We intend to follow the "pseudonym" anonymization method discussed in Section 3, by which personal identifiers are replaced with pseudonyms (or made-up numbers in the case of numerical information, such as particular dates or mentions of one's age), therefore retaining all grammatical information while protecting anonymity. Annotators will carefully mark the linguistic properties, morphological and syntactic information of each phrase to be anonymized and choose a random pseudonym that features the same linguistic properties. To illustrate our anonymization process, consider the fragment sentence in (3), in which a participant's enrollment to a local Icelandic school is discussed:

(3) *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 (*note: this is a school in Reykjavík*).'

The example in (3) contains possibly identifiable information, i.e. the name of a specific school. This sentence could for example be published as in (4), where the original school name has been replaced with a pseudonym -a fake school name-, maintaining the grammatical properties of the original sentence.

(4) *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 does not exist*).'

Similarly, examples (1) and (2) from Section 2 could be anonymized as (5) and (6) respectively:

(5)
Ég var alin upp á Ólafsvík hjá foreldrum mínum, Bjarna og Önnu.
I was raised.FEM in Ólafsvík at parents.DAT mine.DAT, Bjarni.DAT and Anna.DAT
'I was raised in Ólafsvík at my parents, Bjarni and Anna.'

(6)
og pabbi minn var bæjarstjórinn í bænum og mamma vann í móttökunni hjá honum.
and dad.NOM mine.NOM was mayor-the.NOM in town-the.DAT and mom.NOM worked in recepti
'and my dad was the mayor and my mom worked in his reception.'

In addition to replacing the town name "Hólmavík" and the personal names of the parents, "Árni" and "Jóna", with lexically similar names, the occupation "doctor" has been substituted for "mayor". This preserves the grammatical properties of the original utterance and maintains semantic cohesiveness while avoiding possible de-identification of the participant.

References

Adams, A., Aili, E., Aioanei, D., Jonsson, R., Mickelsson, L., Mikmekova, D., Roberts, F., Valencia, J. F., & Wechsler, R. (2019, September). AnonyMate: A toolkit for anonymizing unstructured chat data. In L. Ahrenberg & B. Megyesi (Eds.), *Proceedings of the workshop on nlp and pseudonymisation* (pp. 1–7). Linköping Electronic Press. https://aclanthology.org/W19-6501

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.

Barkarson, S., Steingrímsson, S., & Hafsteinsdóttir, H. (2022). Evolving large text corpora: Four versions of the icelandic gigaword corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2371–2381.

Callegari, E., Nowenstein, I. E., Kristjánsdóttir, I. J., & Ingason, A. K. (2024). Automatic extraction of language-specific biomarkers of healthy aging in icelandic. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1915–1924.

Callegari, E., Sólmundsdóttir, A., & Ingason, A. K. (2023). The acode project: Creating a dementia corpus for icelandic. *CLARIN Annual Conference Proceedings*, 100.

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

Mozes, M., & Kleinberg, B. (2021, March). *No intruder, no validity: Evaluation criteria for privacy-preserving text anonymization*.

Strathern, W., Issig, M., Mozygemba, K., & Pfeffer, J. (2020). *Qualianon - the qualiservice tool for anonymizing text data* (tech. rep. No. TUM-I2087).

CLARIAH-EUS: A Strategic Network Helping Basque Country Researchers to Participate in European Research Infrastructures

Jon Alkorta, Aritz Farwell, Joseba Fernandez de Landa, Begoña Altuna,
Ainara Estarrona, Mikel Iruskieta, Xabier Arregi, Xabier Goenaga,
Jose Mari Arriola
HiTZ centre - Ixa
University of the Basque Country, Spain
jon.alkorta,aritz.farwell,joseba.fernandezdelanda,begona.altuna,
ainara.estarrona,mikel.iruskieta,xabier.arregi,xabier.goenaga,
josemaria.arriola@ehu.eus

Inma Hernáez
HiTZ centre-Aholab
University of the Basque Country
Spain
inma.hernaez@ehu.eus

David Lindemann
Diachronic Linguistics, Typology, and
the History of Basque Research Group
University of the Basque Country
Spain
david.lindemann@ehu.eus

Abstract

CLARIAH-EUS is a node within CLARIAH-ES, Spain's decentralized infrastructure for CLARIN and DARIAH, Europe's leading digital research infrastructures for the humanities, arts, and social sciences. Focusing on Basque or Basque culture-related research in these fields, CLARIAH-EUS offers scholars digital tools and resources. Distinct from other nodes, CLARIAH-EUS serves a language (Basque) rather than a specific territory, making the infrastructure transnational. This article outlines the rationale behind establishing CLARIAH-EUS, its development process, ongoing projects, and future plans

1 Introduction

In the past two decades, a "digital turn" has led to new modes of research and lines of inquiry, with digital tools and methods reshaping the humanities, arts, and social sciences (Arzoz, 2015; Crawford et al., 2014; Terras, 2011). Language technology specifically crafted for these fields is often at the core of this innovative work. Unfortunately, much of this technology is designed for use with English and the development of language technologies is crucial for all languages, especially those spoken by smaller populations (Arzoz, 2015). Basque, one of these languages, has made significant progress in language technology (Gonzalez-Dios & Altuna, 2022; Sarasola et al., 2023) due to collaborative efforts between research groups, foundations, industry clusters, and regional institutions, but it still faces challenges in terms of research maturity and readiness.

The CLARIAH-EUS consortium was established to overcome existing limitations. It seeks to encourage the use of language technology among researchers in Basque-related humanities, arts, and social sciences while also strengthening and facilitating collaboration between these researchers, enabling them to share ideas and innovative approaches more effectively. Consequently, CLARIAH-EUS focuses on language rather than geographical boundaries, making it transnational in scope. Structurally, it functions as a node within CLARIAH-ES, Spain's decentralized infrastructure for CLARIN ERIC and DARIAH ERIC, the two leading digital research infrastructures in Europe for the humanities, arts, and social sciences.

2 Objectives

As mentioned above, one of CLARIAH-EUS's objectives is to support language technology for Basque humanities, arts, and social sciences research. This translates into two key areas. The first is to build a repository that contains digital resources specifically for Basque. These resources will be integrated into the wider CLARIN and DARIAH infrastructures, ensuring that Basque-focused tools are developed and become readily accessible to researchers. The second is to empower researchers by offering them dedicated services and training. We plan to provide the users who are either creating or utilizing Basque language technology for their projects with the resources to work autonomously in the digital domain.

By cultivating these two areas, CLARIAH-EUS wishes to foment a vibrant research community that is dedicated to advancing Basque language technology for the humanities, arts, and social sciences. This focus on collaboration has a twofold purpose. On the one hand, CLARIAH-EUS hopes to open doors to greater participation in international projects. Shared expertise can lead to more impactful outcomes. On the other, the network expects this close collaboration will nourish an environment that sparks groundbreaking approaches to Basque digital humanities and language technology.

3 Funding

CLARIAH-EUS prioritizes securing long-term financial backing. This ensures the viability of research initiatives across the short-, medium-, and long-term. Additionally, it guarantees the ongoing usability and value of the resources created. This focus on sustainability has resonated with several public funding bodies, who have pledged support to the infrastructure. Currently, CLARIAH-EUS is supported by the Basque Government through its Department of Culture and Linguistic Policy,[1] the Provincial Council of Gipuzkoa,[2] and the University of the Basque Country (UPV/EHU). The UPV/EHU is represented by the Vice-Rectorate of Basque, Culture and Internationalization,[3] and by HiTZ, the Basque Center for Language Technology.[4] HiTZ, in addition to providing financial backing, also houses the infrastructure's administrative office. Furthermore, several of its members sit on the CLARIAH-EUS steering committee, contributing their guidance and expertise. Thanks to the support of these institutions, CLARIAH-EUS has assembled a team of four staff members. These individuals play a crucial role, ensuring the smooth operation of both the CLARIAH-EUS infrastructure and the shared administrative office with CLARIAH-ES (also overseen by HiTZ).

4 Evolution of CLARIAH-EUS

The evolution of CLARIAH-EUS has included a design phase (2021-2023) (see sections 4.1. and 4.2) and an implementation phase (2023-present) (see section 4.3).

4.1 First workshop: needs and manifesto

CLARIAH-EUS's first workshop,[5] *Euskararentzako hizkuntza-teknologia Humanitateetan eta Zientzia Sozialetan garatzeko CLARIAH-EUS azpiegitura diseinatzen* (*Designing the CLARIAH-EUS infrastructure to develop language technology for Basque in the Humanities and Social Sciences*), organized by HiTZ on November 26, 2021, lay the foundation for the future infrastructure.

The workshop aimed to foster discussion about opportunities and needs across various research areas related to Basque language and culture, featuring several activities: 1) the presentation of a collection of use cases and posters depicting digital projects relevant to Basque studies, which provided a platform for researchers to share their work; 2) collaborative breakout sessions focused on identifying the strategic resources most crucial for Basque research across different disciplines; and 3) engaging and building bridges between researchers that encouraged active participation in CLARIAH-EUS's future.

[1] https://www.euskadi.eus/eusko-jaurlaritza/kultura-hizkuntza-politika-saila/

[2] https://www.gipuzkoa.eus/eu/

[3] https://www.ehu.eus/eu/web/nazioarteko-harremanak

[4] https://www.hitz.eus/eu

[5] https://www.clariah.eus/eu/1-workshop

The event drew participation from nine institutions and thirty-four researchers representing twenty distinct research groups. Fourteen projects were presented and 134 organizations and individuals signed a manifesto.[6] This collective voice underscored the widespread demand for a dedicated digital humanities infrastructure for Basque research.

4.2 Assembling the network

Between 2021 and 2023, CLARIAH-EUS's goal was to pursue the backing of several organizations and research groups. This was procured from ten entities: HiTZ (UPV/EHU), Udako Euskal Unibertsitatea (UEU), Iker research group, Elhuyar, Gogo Elebiduna research group (UPV/EHU), Elebilab research group (UPV/EHU), Aholab research group (UPV/EHU), Ixa research group (UPV/EHU), Soziolinguistika Klusterra, and the Unesco Chair in Human Rights and Public Powers (UPV/EHU). Nine of these institutions and research groups hail from the southern Basque Country and one (Iker) is from the northern Basque Country. During this time, CLARIAH-EUS's position within CLARIAH-ES in Spain and CLARIN and DARIAH at the European level was further solidified.

4.3 Second workshop: community and organization

CLARIAH-EUS held its second workshop[7] in November 2023, in which we presented the CLARIAH-EUS infrastructure and existing Basque digital humanities projects. The workshop marked a significant milestone in the form of a kickoff ceremony for the founding members. The focus shifted from initial brainstorming to outlining the infrastructure's future. Key discussions centered on CLARIAH-EUS's organizational structure and a road map, which put an emphasis on strategic directions for the next five years. Two invited speakers shared their expertise and twenty-one research groups presented posters. A selection of these, along with details about the participating research groups, will be featured in a forthcoming publication, offering a valuable glimpse into the Basque digital humanities landscape.[8]

5 Resources

As highlighted above, a core objective of CLARIAH-EUS is to empower researchers with the tools and resources[9] they need to excel in digital humanities and social sciences. These resources fall into two categories: 1) pre-existing resources that existed before CLARIAH-EUS was established. The infrastructure has integrated these into the network to maximize their reach and usability for the research community; and 2) newly developed resources that CLARIAH-EUS has created. The following includes examples from both categories: 1) Parlamint-ES-PV 4.0 (Alkorta & Iruskieta, 2022); 2) four datasets for Computational Social Science (Fernandez de Landa et al., 2019), (Fernandez de Landa & Agerri, 2021), (Agerri et al., 2021) and (Fernandez de Landa et al., 2024); and 3) BIM (*Basque in the Making (BIM): A Historical Look at a European Language Isolate*) and SAHCOBA (*Syntactically Annotated Historical Corpus in Basque*) projects (Estarrona et al., 2022).

6 Subsequent Steps

In the near future, CLARIAH-EUS aims to add a CLARIN B-centre to its CLARIN K-centre. This will allow us to offer technical services as well as valuable instructional guidance to researchers. Three key criteria will guide CLARIAH-EUS's development:

- **Building Resources**. CLARIAH-EUS will prioritize creating or adapting resources and services that are readily accessible to researchers through the CLARIAH-EUS node.
- **Strategic Focus**. The infrastructure will target resources and services that strategically address the needs of the Basque research community.

[6] https://www.clariah.eus/eu/manifestua

[7] https://www.donostiakultura.eus/eu/ikastaroak/clariah-eus-euskararako-ikerketa-azpiegitura-eraikitzen

[8] https://www.clariah.eus/eu/2-workshopa-azpiegitura-eraikitzen

[9] https://www.clariah.eus/eu/baliabideak_sailkapena

- **Collaboration**. CLARIAH-EUS will create or adapt resources and services that seamlessly integrated with CLARIN and DARIAH.

CLARIAH-EUS is focusing on making an immediate impact by adapting existing resources and incorporating them into CLARIN and DARIAH. Specifically, we hope to include the Analhitza tool (Otegi et al., 2017), the Euscrawl system (Artetxe et al., 2022), and ParlaMint. Additionally, our plan involves providing various corpora, including literature, historical texts, and social network data, alongside developing new resources like a data repository and APIs. Looking ahead, we aim to create tools and resources for sociology, journalism, literature, and history, ideally aligning with GLAM-related initiatives.

Acknowledgments

We thank the Basque Government and its Department of Culture and Linguistic Policy, the Provincial Council of Gipuzkoa, the Vice-Rectorate of Basque, Culture and Internationalization at the University of the Basque Country (UPV/EHU), and the HiTZ center for their support, as well as the CLARIAH-ES Network (RED2022-134527-E) of the Spanish Ministry of Science, Innovation and Universities.

References

Agerri, R., Centeno, R., Espinosa, M., Fernandez de Landa, J., & Rodrigo, A. (2021). VaxxStance@ IberLEF 2021: overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, *67*, 173–181.

Alkorta, J., & Iruskieta, M. (2022). Adding the Basque Parliament Corpus to ParlaMint Project. *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, 107–110.

Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O., & Soroa, A. (2022). Does Corpus Quality Really Matter for Low-Resource Languages? In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 7383–7390).

Arzoz, X. (2015). The Impact of Language Policy on Language Revitalization: The Case of the Basque Language. *Cultural and Linguistic Minorities in the Russian Federation and the European Union: Comparative Studies on Equality and Diversity*, 315–334.

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication*, *8*(0), 1663–1672.

Estarrona, A., Etxeberria, I., Soraluze, A., Etxepare, R., & Padilla-Moyano, M. (2022). The first annotated corpus of historical Basque. *Digital Scholarship in the Humanities*, *37*(2), 391–404.

Fernandez de Landa, J., & Agerri, R. (2021). Social analysis of young Basque-speaking communities in twitter. *Journal of Multilingual and Multicultural Development*, *0*(0), 1–15.

Fernandez de Landa, J., Agerri, R., & Alegria, I. (2019). Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. *Information*, *10*(6).

Fernandez de Landa, J., García-Ferrero, I., Salaberria, A., & Campos, J. A. (2024). Uncovering social changes of the basque speaking twitter community during covid-19 pandemic. *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, 363–371.

Gonzalez-Dios, I., & Altuna, B. (2022). Natural Language Processing and Language Technologies for the Basque Language. *Cuadernos Europeos de Deusto*, (04), 203–230.

Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., & Uria, L. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research.

Sarasola, K., Aldabe, I., Díaz de Ilarraza, A., Estarrona, A., Farwell, A., Hernáez, I., & Navas, E. (2023). Language Report Basque. In *European language equality: A strategic agenda for digital language equality* (pp. 95–98). Springer.

Terras, M. (2011). Quantifying digital humanities. *UCL Centre for Digital Humanities*.

Enriching the ParlaMint-DK corpus with Policy Domains

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

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

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

Abstract

In this paper, we present ParlaMint-DK 4.1 containing the Danish parliamentary speeches from 2014 to 2022 enriched with the annotation of policy domains. The policy domains were semi-automatically added to the speeches using the agenda items of the meetings in which the speeches occurred. Moreover, in this version of the corpus, we corrected some of the lemma and NER errors occurring in the previous ParlaMint-DK version by improving the mapping between the tools used in the linguistic annotation process. In the paper, we also present a first analysis of the distribution of the various policy domains in the speeches.

1 Introduction

In this paper, we present the newly published version of the ParlaMint-DK (4.1) containing policy areas annotations and improved linguistic annotations. ParlaMint-DK is the Danish part of the corpora collected and annotated under the ParlaMint project (Erjavec, Kopp, Ogrodniczuk, Osenova, Agerri, et al., 2024) and (Erjavec, Kopp, Ogrodniczuk, Osenova, Agirrezabal, et al., 2024). The annotation of policy areas or domains in political speeches and other political material has been addressed by many researchers, e.g., (Baumgartner et al., 2011; Ristilä & Elo, 2023; Yu et al., 2023; Zirn et al., 2016), since the identification of the policy areas in the data facilitates the study of how politicians from different political wings, parties and countries have addressed them.

We started annotating policy areas in the Danish parliamentary speeches deposited in CLARIN-DK in a pilot study described in (Hansen et al., 2019). The speech corpus annotated with policy areas from October 2009 to June 2017 was released in 2021 as the Danish Parliament corpus with subject annotations v.2 (Hansen & Navarretta, 2021). This corpus, as the ParlaMint-DK corpus, was downloaded from the Danish Parliament (*Folketinget*)'s website ftp://oda.ft.dk. It was released in CSV format and each speech was annotated with one or two policy domains. The corpus also contains information about the speakers, their party, age and gender. Classification algorithms were run on the corpus to test how well they identified the speeches' main policy domain Hansen et al. (2019) and the primary and secondary policy domains in those speeches that are annotated with two domains (Navarretta & Hansen, 2022). The policy domains annotations of this corpus have been used in a number of studies, e.g., for the investigation of how *Immigration* and *Environment* have been dealt with by different parties in the covered period (Navarretta & Hansen, 2023; Navarretta et al., 2022) .

We have recently coded the policy domains of the speeches from June 2017 to June 2022 and added the policy domains annotations to all the speeches in the ParlaMint-DK corpus.

In what follows, we shortly describe background work about the classification of policy areas (section 2), while in section 3 we account for the classification system, which we adopted, and the semi-automatic annotation method which we applied. In section 4, we describe the new version of the ParlaMint-DK corpus, in which also some of the linguistic annotations have been improved, and we present the distribution of the various policy domains in the corpus. Finally in section 5, we conclude and discuss future work.

2 Background work

Various policy domain classification systems have been proposed for different purposes. The most known classification systems are the following: a) the system adopted by the Comparative Manifesto Project (CMP)[1], and b) the one proposed by the Comparative Agendas Project (CAP)[2].

The CMP classification is fine-grained and comprises more than 550 categories used to annotate so called quasi-sentences in numerous party election manifestos from many countries. This classification allows the analysis of policy preferences expressed by political parties in their manifestos.

The scheme used in the Comparative Agendas Project (CAP) builds upon the classification applied in the Policy Agendas Project[3], who aimed to structure the U.S. policy data. The CAP scheme is a modified version of this classification, and it aims to account not only for the policy activities of the U.S. data, but also for those of other countries all over the world Baumgartner et al., 2011. The CAP classification scheme comprises 21 main domain categories and 192 sub-categories. Danish researchers in political science from the University of Aarhus have manually annotated political data from 1953 to 2007 in the Danish Policy Agendas Project[4] using an adapted version of the CAP scheme relating policy activities in the parliamentary debates to the CAP classes taking into account the responsibility areas of the Danish Parliament's committees. We followed their suggestion as describe in (Hansen et al., 2019).

3 The classification of policy areas

The classification scheme of policy domains, which we have applied in ParlaMint-DK 4.1, consists of 20 classes. 19 of them correspond to the areas of responsibilities in the Danish Parliament (spokesmanships) in the covered period, while one class *Other* covers government operations and issues that do not fall under CAP. Table 3 shows the 20 policy domain classes, the corresponding areas of responsibility in the Danish parliament, the corresponding CAP codes and CAP areas.

3.1 The annotation method

The policy domains annotations were semi-automatically added to the Danish speeches in The Danish Parliament Corpus (2009-2017) extracting them from the titles of the agenda items of the meetings. The method was described in (Hansen et al., 2019) where the first pilot annotation of part of the corpus was presented. The method consists of the following steps: 1) extraction of the agenda titles 2) normalization, e.g., "Third reading of bill N: XYZ" becomes ”XYZ", 3) manual annotation of the agenda titles with one or two policy areas, and 4) automatic assignment of the policy area(s) to each speech under the meeting covered by the relevant agenda titles. The speeches and the annotations were in CSV format and the policy domains were added to the existing annotations in excel. 5000 speeches which were coded with two policy areas were reviewed by two annotators independently also in excel. The two annotators, did not find any errors in the assignment of the two policy areas, but in a few cases they disagreed on which of the two annotated areas should be considered the primary one (Hansen et al., 2019).

The extended annotations of policy domains covering the speeches from 2017 to 2022 were performed according to the same methodology as in Hansen et al., 2019, but, in the annotations of policy domains in the ParlaMint-DK corpus, we also decided to use an extra domain *Other* accounting for the speeches about government operations and other issues. To add all the annotations to the ParlaMint-DK speeches, we first created a TEI taxonomy over the policy domains, and then the policy domain information was added to each speech as an *@ana* attribute in the *u*-element[5]. Since all the speeches from the Danish Parliament have a unique identifier, this step was trivial.

[1] https://manifesto-project.wzb.eu/

[2] https://www.comparativeagendas.net/

[3] https://liberalarts.utexas.edu/government/news/feature-archive/the-policy-agendas-project.php

[4] http://www.agendasetting.dk/.

[5] This was done after consulting Tomaž Erjavec.

Policy Domain	Area of Responsibility	CAP no.	CAP Areas
Economy	Finance, Fiscal Affairs	1	Domestic Macroeconomic Issues
Health Care	Psychiatry, Health	3	Health
Agriculture	Animal Welfare, Fisheries, Food, Agriculture	4	Agriculture
	Consumer Policy	1525	Consumer Policy
Labour	Labour market	5	Labour and Employment
Education	Higher Education and Research Education	6	Education
Environment	Environment	7	Environment
Energy	Energy	8	Energy
	Climate	705	Air and Noise Pollution, Climate Change and Climate Policies
Immigration	Immigration and Integration, Alien Affairs, Naturalization	9	Immigration and Refugee Issues
Infrastructure	Transportation	10	Transportation
	IT, Media	17	Space, Science, Technology and Communications
Justice	Legal Affairs	12	Law, Crime, and Family Issues
	Constitutional Matters	20	Government Issues
Social Affairs	Children, Family, Social Services, Senior Citizens	13	Social Welfare
	Gender Equality	2	Civil Rights, Minority Issues, and Civil Liberties
Housing	Housing	14	Community Development & Housing Issues
Local and	Rural Districts and Islands	4	Community Development & Housing Issues
Regional Affairs	Municipal Affairs	2001	Local Government Issues
Business	Trade and Industry	15	Industrial and Commercial Policy
Defence	Defence	16	Defence
Foreign Affairs	Foreign Affairs, Development, Cooperation	19	International Affairs and Foreign Aid
European Integration	EU	1910	International Affairs and Foreign Aid
Territories	Faroe Islands, Greenland	2105	Dependencies and Territorial Issues
Culture	Cultural Affairs	23	Cultural Policy Issues
	Ecclesiastical Affairs	210	The Danish national church
	Sport	1526	Sport and Gambling
Other	-	-	-

Table 1: Policy domains, and corresponding responsibility areas, CAP numbers, and CAP areas in ParlaMint-DK

4 ParlaMint-DK 4.1

The linguistic annotations were performed as in the former versions of the corpus via the Text Tensorium, using ten different tools in a workflow that comprises twelve steps. Evaluating the linguistic annotations of part of ParlaMint-DK 4.0, we found some systematic lemma annotation errors, and we avoided them by taking morphology and word form into account when mapping between the Universal tag set output by UD-pipe and the CST tag set used by CSTlemma. Already for the first published version of the ParlaMint-DK corpus we decided not to use the UD-pipe software for delivering lemma annotations because, as a lemmatiser, UD-pipe performs worse than CSTlemma. Also, we required that the application of a lemmatization rule was conditioned on the word class of the input word, while UD-pipe often applies lemmatization rules that are not meant for the word classes assigned by UD-pipe. Previously, the mapping from the UD-tag set onto the CST tag set was performed by CSTlemma itself, using a simple lookup table. Now, the mapping task is delegated to a separate tool, the PoS translator. The tool combines information from tokenization, PoS-tagging and morphological analysis to provide the correct information to CSTlemma. Also the NER annotations were improved in this version of the annotated corpus. This improvement especially dealt with the abbreviations of parties' and organisations' names.

ParlaMint-DK 4.1 comprises 398,610 speeches. After having removed utterances made by the Chair (the Speaker)[6], and speeches which did not address a policy domain[7], we have 208,881 speeches with policy domains. Figure 1 shows the distribution of the main policy domains in the corpus. The most frequently addressed policy domain in the speeches is *Economy*, followed by *Immigration*, *Justice*, and *Other*. The prominence of the *Economy* domain is not surprising, while the frequency of speeches about *Immigration* indicates the importance of this topic in Danish politics the past ten years.

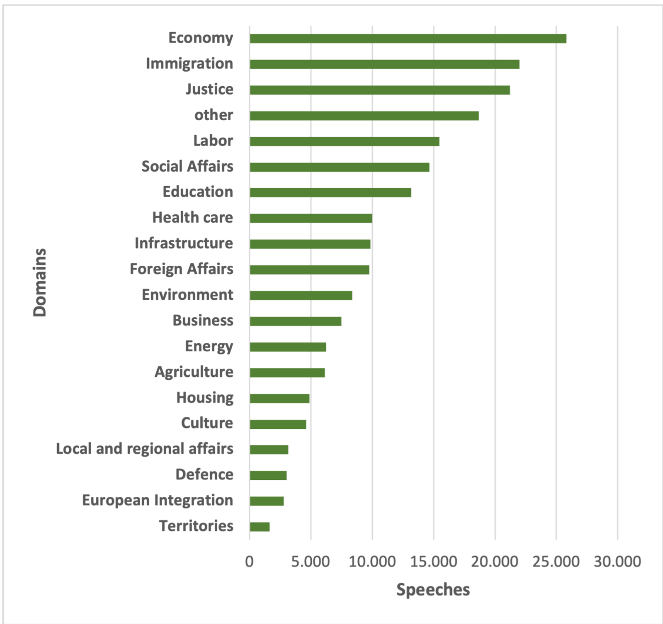


Figure 1: The distribution of the main policy domains in ParlaMint-DK

5 Conclusions and Future Work

In the paper, we have presented the forthcoming version of ParlaMint-DK with improved lemma and NER annotations and policy domain coding. We are now conducting studies on these data such as investigations of the mostly frequently addressed domains by female and male politicians and by various parties.

References

Baumgartner, F. R., Jones, B. D., & Wilkerson, J. (2011). Comparative Studies of Policy Dynamics. *Comparative Political Studies*, *44*(8), 947–972. https://doi.org/10.1177/0010414011405160

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agerri, R., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., … Fišer, D. (2024). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1911

[6] The utterances made by the Chair have the same policy domain annotation as the speeches that are chaired in that section, if they occur under an agenda point with that policy domain. They should not been considered when analysing the content of policy domains, and therefore we have removed them in our analyses. However, these utterances can be interesting when accounting for the dynamics of the debates.

[7] This speeches have no domain annotation.

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., Aires, J., Albini, M., Alkorta, J., Antiba-Cartazo, I., Arrieta, E., Barcala, M., Bardanca, D., Barkarson, S., Bartolini, R., Battistoni, R., Bel, N., Bonet Ramos, M. d. M., Calzada Pérez, M., … Fišer, D. (2024). Multilingual comparable corpora of parliamentary debates ParlaMint 4.1 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1912

Hansen, D. H., & Navarretta, C. (2021). The danish parliament corpus 2009 - 2017, v2, w. subject annotation [CLARIN-DK-UCPH Centre Repository]. http://hdl.handle.net/20.500.12115/44

Hansen, D., Navarretta, C., Offersgaard, L., & Wedekind, J. (2019). Towards the Automatic Classification of Speech Subjects in the Danish Parliament Corpus. *CEUR Workshop Proceedings*, *2364*, 166–174.

Navarretta, C., Haltrup Hansen, D., & Jongejan, B. (2022, June). Immigration in the manifestos and parliament speeches of Danish left and right wing parties between 2009 and 2020. In D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the LREC22 Workshop ParlaCLARIN III* (pp. 71–80). ELRA. https://aclanthology.org/2022.parlaclarin-1.11

Navarretta, C., & Hansen, D. H. (2023, September). According to BERTopic, what do Danish parties debate on when they address energy and environment? In C. Klamm, G. Lapesa, V. Gold, T. Gessler, & S. P. Ponzetto (Eds.), *Proceedings of the 3rd KONVENS Workshop on Computational Linguistics for the Political and Social Sciences* (pp. 59–68). Association for Computational Lingustics. https://aclanthology.org/2023.cpss-1.6

Navarretta, C., & Hansen, D. H. (2022). The Subject Annotations of the Danish Parliament Corpus (2009-2017) - Evaluated with Automatic Multi-label Classification. *Proceedings of LREC 2022*.

Ristilä, A., & Elo, K. (2023). Observing political and societal changes in Finnish parliamentary speech data, 1980–2010, with topic modelling. *Parliaments, Estates and Representation*, 1–28.

Yu, H.-C., Rehbein, I., & Ponzetto, S. P. (2023). Policy domain prediction from party manifestos with adapters and knowledge enhanced transformers. *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, 229–244.

Zirn, C., Glavas, G., Nanni, F., Eichorst, J., & Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016)*, 88–93.

Expansion of the RomCro corpus with texts in Catalan

Bojana Mikelenić
University of Zagreb, Croatia
bmikelen@ffzg.unizg.hr

Antoni Oliver
Open University of Catalonia, Barcelona, Spain
aoliverg@uoc.edu

Marko Tadić
University of Zagreb, Croatia
mtadic@ffzg.unizg.hr

Abstract

In this article we describe an existing multilingual and multidirectional parallel corpus RomCro and its expansion with texts in Catalan. This corpus is composed of literary texts in five Romance languages (Spanish, French, Italian, Portuguese, Romanian) and Croatian, with a total of 142K segments and 15.7 Mw. In this first expansion phase, we will add 16 Catalan translations, to be followed later by three originals in Catalan and their translations into the other languages. The corpus will be available on the HR-CLARIN repository.

1 Introduction

The construction of the corpus RomCro (Bikić-Carić et al., 2023) began in autumn 2019 and is funded by the Faculty of Humanities and Social Sciences of the University of Zagreb. RomCro is a multilingual and multidirectional Romance-Croatian parallel corpus. It contains original literary texts in six languages (Spanish, French, Italian, Portuguese, Romanian and Croatian), as well as their translations into the other five languages. The corpus consists of 142,470 Translation Units (TUs) and 15.7 million words. The distribution by language is as follows: French 2.8 Mw, Spanish 2.7 Mw, Romanian 2.6 Mw, Italian 2.6 Mw, Portuguese 2.6 Mw, and Croatian 2.4 Mw.

In this paper, we will first contextualize the corpus comparing it to other similar resources, then detail its building and, finally, explain the process of its expansion with Catalan translations.

2 Previous work

The majority of available multilingual parallel corpora tend to be automatically processed and rarely include literary texts.[1] For example, very large parallel corpora, such as CCMatrix[2] (Schwenk et al., 2021) and MultiCCAligned[3] (El-Kishky et al., 2020) contain segmentation and alignment errors, while corpus such as OpenSubtitles[4] (Tiedemann, 2012; Lison & Tiedemann, 2016) are limited with the specific format of subtitles, seldom unknown source language and presents quite noisy source. On the other hand, there are language resources based on vast legal and other documents of the European Union, for example JRC-Acquis (Steinberger et al., 2006), that contain very specific vocabulary and are mostly translated from two or three languages to all the rest, and regular updates of TMs from DGT (Steinberger et al., 2013).

Literary corpora are specific because of the legal issues concerning the copyrighted material and because they usually require a lot of manual work.[5] Even so, there are some multilingual literary corpora

[1] For a list of bilingual and multilingual parallel corpora available through the CLARIN infrastructure please see: https://www.clarin.eu/resource-families/parallel-corpora

[2] https://github.com/facebookresearch/LASER/tree/main/ tasks/CCMatrix

[3] https://www.statmt.org/cc-aligned/

[4] https://www.opensubtitles.org/

[5] A list of literary corpora, monolingual and multilingual, that can be accessed via CLARIN, can be found here: https://www.clarin.eu/resource-families/literary-corpora

including most or all of the languages in RomCro, for example TransLiTex (Fraisse et al., 2018) or the literary subcorpus in InterCorp (Čermák, 2019). Nevertheless, TransLiTex contains translations of one book, Mark Twain's *Adventures of Huckleberry Finn*, into 23 languages, so the data size for each language is considerably smaller in comparison to RomCro, and InterCorp comprises 40 languages, with Czech serving as the pivot language. In that sense, neither of these corpora are completely multidirectional.

RomCro is a different corpus from those already available for several reasons. It is a literary corpus of 15.7 Mw of published originals and translations, which ensures a high quality of the data. Even though the initial processing of the texts and the alignment were done automatically, the results were later reviewed and corrected by hand, thus creating a more reliable resource. Finally, it is almost completely multidirectional, making the source language always known and including both original and translated material in each of the six languages.

3 RomCro – Stages of development

The corpus building took three years, with some texts added later, but the first version of RomCro was published in 2023. The process involved six stages: 1) Selection and collection of texts, 2) Digitization of texts, 3) Preparation for segmentation and sentence alignment, 4) Segmentation, alignment, and manual correction, 5) Lemmatization and morphosyntactic annotation, and 6) Access to the corpus.

When considering the corpus outline and selecting the texts, the following criteria were adopted: availability, quality, synchronicity and linguistic homogeneity. A primary challenge was sourcing high-quality translations from the original language into other languages, leading to the choice of literary texts. The unequal distribution of originals in each language (Table 1) reflects the higher availability of titles translated from certain languages (e.g., Spanish) compared to others (e.g., Croatian).

To maintain synchronicity, texts had to be relatively recent publications, posing difficulty for languages like Croatian and, more notably, Romanian, where two titles from the first half of the 20th century were selected. Additionally, having in mind the complexity of this resource, the availability of material was given primacy over equal distribution, so for some languages more texts first published in the past century were included in the corpus (e.g., originals in French).

For linguistic homogeneity, only European varieties of Spanish, Portuguese and French were meant to be included. However, since four titles were available only in Brazilian Portuguese,[6] they were added to the corpus with the option to exclude them when filtering by provided notes.

Digitization involved scanning texts and utilizing Optical Character Recognition via Abbyy FineReader[7]. Manual correction in MS Word prepared the material for segmentation and alignment, with the help of undergraduate and master's level students collaborating on the project.

Segmentation and alignment were completed using LF Aligner,[8] a freely available tool based on Hunalign (Varga et al., 2005), with subsequent manual revision. Lemmatization and morphosyntactic annotation employed FreeLing (Padró, 2011) for Spanish, French, Italian, and Portuguese, and MULTEXT-East (Erjavec et al., 2003; Erjavec, 2017) for Romanian and Croatian.

The number of original titles included in the corpus is 27, so adding the translations to all other languages brings the total to 162 texts. However, there are three translated texts that are not yet available,[9] so the corpus currently counts with 159 texts. As illustrated in Table 1, there are seven originals in Spanish (es), six in French (fr), four in Italian (it), four in Romanian (ro), three in Portuguese (pt), and three in Croatian (hr).

The lemmatized and POS tagged version is accessible on Sketch Engine (Kilgarriff et al., 2004), while the untagged TMX and TSV versions can be found on the ELRC platform[10] under the CC-BY-NC-4.0 license. In both formats, the order of languages is as follows: Spanish (es), French (fr), Italian (it), Portuguese (pt), Romanian (ro), Croatian (hr). It is important to mention that the corpus available on

[6] These Portuguese translations are: *A fada carabina* by Daniel Pennac, *A forma da água* by Andrea Camilleri, *Acontecimentos na Irrealidade Imediata* by Max Blecher, and *Nostalgia* by Mircea Cărtărescu.

[7] https://pdf.abbyy.com/

[8] https://sourceforge.net/projects/aligner/

[9] The book *Dora i Minotaur: Moj život s Picassom* is still not translated from Croatian to Spanish and Portuguese, while *El asombroso viaje de Pomponio Flato* is not available in Romanian.

[10] https://elrc-share.eu/repository/search/?q=romcro

this platform is an earlier version and is missing two texts.[11] All versions contain notes on the original language, writer, and title, with segment order scrambled to protect copyright.

		Titles:
1	ES	La sombra del viento (C.R. Zafón, 2001)
2		La catedral del mar (I. Falcones, 2006)
3		El juego del ángel (C.R. Zafón, 2008)
4		El asombroso viaje de Pomponio Flato (E. Mendoza, 2008)
5		Soldados de Salamina (J. Cercas, 2001)
6		El mapa del tiempo (F. J. Palma, 2008)
7		El tiempo entre costuras (M. Dueñas, 2009)
8	FR	Seras-tu là ? (G. Musso, 2006)
9		HHhH (L. Binet, 2010)
10		Un barrage contre le Pacifique (M. Duras, 1950)
11		La Fée Carabine (D. Pennac, 1987)
12		L'amant (M. Duras, 1984)
13		A l'ombre des jeunes filles en fleur (M. Proust, 1919)
14	IT	Imprimatur (Monaldi & Sorti, 2002)
15		Le otto montagne (P. Cognetti, 2017)
16		La forma dell'acqua (A. Camilleri, 1994)
17		L'amica geniale (E. Ferrante, 2011)
18	RO	Maitreyi (M. Eliade, 1933)
19		Întâmplări în irealitatea imediată (M. Blecher, 1936)
20		Nostalgia (M. Cărtărescu, 1993)
21		Cartea şoaptelor (V. Vosganian, 2009)
22	PT	A viagem do elefante (J. Saramago, 2008)
23		Nenhum olhar (J. L. Peixoto, 2000)
24		As intermitências da morte (J. Saramago, 2005)
25	HR	Muzej bezuvjetne predaje (D. Ugrešić, 1998)
26		Mediteranski brevijar (P. Matvejević, 1987)
27		Dora i Minotaur: Moj život s Picassom (S. Drakulić, 2015)

Table 1. Originals in each language included in the corpus

The corpus is almost completely multidirectional, so each language serves as a source and a target language. The highest percentage of original texts is in Spanish, followed by French, Italian, Romanian, Portuguese, and Croatian, while the percentage of translated texts follows the reverse pattern, being lowest in Spanish and highest in Croatian. There is currently work underway to add more Portuguese and Croatian originals, since these have the highest number of translated segments.

4 Expansion with translations to Catalan

In the following months, we plan to incorporate Catalan into the RomCro corpus. This addition will also expand the number of available parallel corpora for Catalan within the CLARIN framework. Currently, CLARIN has only two parallel corpora for Catalan: ParlaMint 4.0 and MaCoCu-ca-es 1.0. To our knowledge, RomCro will also stand as the sole literary parallel corpus accessible for Catalan.

The incorporation of Catalan into the RomCro corpus will be performed in two distinct phases. First, we will include the existing translations of novels already present in RomCro into Catalan. Subsequently, we will add Catalan novels that have been translated into all the other languages.

A preliminary inspection reveals that out of the 27 novels already incorporated into the RomCro corpus, 16 have been translated into Catalan. Additionally, for each source language represented in RomCro—except for Croatian—there exists a translation of at least one novel into Catalan. We have also selected three novels from three Catalan authors translated into all the languages in RomCro: Jaume

[11] These are the Italian translation of Croatian novel *Muzej bezuvjetne predaje* and the Portuguese translation from Romanian of *Maitreyi*.

Cabré – *Les veus del Pamano*; Albert Sánchez Piñol – *La pell freda*; and Mercè Rodoreda – *La Plaça del Diamant*. These will be added later. This updated version, i.e. RomCro v2.0 will be available on the HR-CLARIN repository.

5 Conclusion

We presented a multilingual and multidirectional literary parallel corpus of six languages with a total of 15.7 Mw. The next expansion phase will include adding Catalan originals and translations to the other languages. The corpus is unique for its language combination and for being one of a few almost completely multidirectional literary parallel corpora of this size.

Acknowledgments

This work was supported by the Croatian Science Foundation under the project number MOBODL-2023-08-9511, funded by the European Union – NextGenerationEU.

References

Bikić-Carić, G., Mikelenić, B., & Bezlaj, M. (2023). Construcción del RomCro, un corpus paralelo multilingüe. *Procesamiento del Lenguaje Natural, 70*, 99–110. Sociedad Española para el Procesamiento del Lenguaje Natural.

Čermák, P. (2019). InterCorp. A parallel corpus of 40 languages. In Doval, I., & Sánchez Nieto, M. T. (Eds.), *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*, 93–101. Philadelphia / Amsterdam: John Benjamins Publishing Company.

El-Kishky, A., Chaudhary, V., Guzman, F., & Koehn, P. (2020). CCAligned: A Massive Collection of Cross-lingual Web-Document Pairs. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020),* 5960–5969. Association for Computational Linguistics.

Erjavec, T. (2017). MULTEXT-East. In Ide, N., & Pustejovsky, J. (Eds.), *Handbook of Linguistic Annotation*, 441–462. Springer Dordrecht.

Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., & Vitas, D. (2003). The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, 25–32. ACL (Budapest).

Fraisse, A., Tran, Q.-T., Jenn, R., Paroubek, P., & Fisher Fishkin, S. (2018). TransLiTex: A Parallel Corpus of Translated Literary Texts. *Proceedings of the 11th Language Resources and Evaluation Conference*. European Language Resource Association.

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the Eleventh EURALEX International Congress,* 105–116.

Lison, P., & Tiedemann J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 923–929. European Language Resource Association.

Padró, L. (2011). Analizadores Multilingües en FreeLing. *Linguamatica, 3*(2), 13–20.

Schwenk, H., Wenzek, G., Edunov, S., Grave, É., Joulin, A., & Fan, A. (2021). CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 6490–6500.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2142–2147. European Language Resource Association.

Steinberger, R., A. Eisele, S. Klocek, S. Pilos, & Schlüter, P. (2013). DGT-TM: A freely available translation memory in 22 languages. *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012),* European Language Resource Association.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214–2218.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Nemeth, L., & Tron, V. (2005). Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, 590–596.

Benchmarking and Research Infrastructures: Evaluating Dutch Automatic Speech Recognition

Dragoș Alexandru Bălan
University of Twente
d.a.balan@utwente.nl

Khiet Truong
University of Twente
k.p.truong@utwente.nl

Henk van den Heuvel
Radboud University
henk.vandenheuvel@ru.nl

Roeland Ordelman
University of Twente
roeland.ordelman@utwente.nl

Abstract

In this paper we present the setup and results of a Dutch Open Speech Recognition Benchmark strategy, initiated in the course of use cases in three research infrastructure projects in The Netherlands, to help scholars, infrastructure providers, and speech researchers make informed choices about which speech recognition engines, configurations and models to use, to facilitate or to improve upon. The obtained results are a tangible starting point to expand collaboration and optimize the coverage of the speech types and conditions addressed in the benchmark.

1 Introduction

With the aim to support digital scholarship, the development, preservation, and provisioning of (scientific) software commonly referred to as "tools" has been a central theme for Research Infrastructures (RIs) in the Arts and Humanities for long. Generally, tools are being identified by collecting scholarly use cases or research scenarios and extracting scholarly workflows: sets of infrastructure facilities, resources, and services that together serve the requirements of humanities scholars.By collecting use cases within research disciplines, and translating these to workflows consisting of infrastructure components, RIs are able to identify which workflows need to be facilitated, and to define a catalog of tools for an RI to provide. In order to help scholars discover the tools they need for their specific use scenario, RIs setup *registers* such as the CLARIN Virtual Language Observatory[1] (VLO) and the SSHOC Marketplace[2].

However, "helping scholars find tools they need", should be interpreted as something that requires more than listing the software a community has produced in online databases. Here, perspectives on optimizing *findability* and *reuse* of tools coincide with perspectives on scholarly *training*; not only on how a tool can or should be used properly in a research workflow, but specifically also on raising awareness of the limitations of tools, sometimes referred to as *tool criticism* (Koolen et al., 2018). Findability and reuse perspectives put emphasis on the information components that are needed to enable researchers to make a proper assessment of whether a tool can be attributed to their research or not, such as a tool's Technology Readiness Level (TRL), but also the "user readiness" of tools (for example, how a tool fits in with a research methodology or the digital context a researcher is used to work in).

An information component that is frequently mentioned to be an important element, but not so frequently provided, is *performance*. Implicitly, performance can be derived from reported scientific results accomplished by deploying a tool, or how frequently it is used in a research community. Often, however, explicit information on performance levels is required and asked for. In the case of automatic speech recognition (ASR), being able to assess the quality of the data enrichment process that decodes speech in a researcher's data set (e.g., audiovisual media archives, interviews, conversations) to a textual representation, is crucial for interpreting results from search and quantitative data analysis, as performance levels in ASR can vary drastically depending on the type of speech, acoustic condition, context (e.g., domain vocabulary), and characteristics of the speakers.

[1] https://vlo.clarin.eu/

[2] https://marketplace.sshopencloud.eu/

In providing ASR performance information, the different speech types and/or conditions that could be encountered in research data sets is one aspect, the other is the range of ASR engines, configurations and models that are available, especially since these are subject of rapid changes due to ongoing advances in the field of machine learning. Scholars want to know whether applying automatic speech recognition to their data would be helpful or not, and if so, which engine they should turn to. If performance measurements were below expectations, the scholar could decide to invest in the manual annotation (Gref et al., 2022) of (parts of) the data set, instead of wasting time making sense out of noisy data. But also for a research infrastructure supplier that would facilitate a (possibly large-scale) data enrichment workflow (e.g., data ingest, computational resources), performance levels are important to decide on the ASR tool(s) it should provide in its infrastructure or not. Finally, for researchers interested in improving available ASR (models), performance statistics would help to determine whether an improvement strategy would be feasible, for example given available data sets that can be used for training.

In the context of these stakeholders in need of ASR performance information, we decided to develop a *collaborative benchmarking strategy* that could feed into the general concept of tool criticism in the context of research infrastructures.

2 Experimental Setup

Benchmarking is a widely used practice to critically and quantitatively assess the performance of various (AI) models or algorithms on a specific topic or dataset. Research groups, both in humanities and computer science, are interested in ASR performance on speech types, depending on their specific research focus. Our research aims to create a Dutch Open Speech Recognition Benchmark initiative, providing a matrix of ASR performances relative to speech types. In this paper, we discuss the essential parts of the setup and results and refer to the official benchmark website[3] for more detailed information. The benchmark was initiated in the context of Dutch research infrastructure projects focusing on research use cases on specific speech types: audiovisual media (Ordelman et al., 2018), Oral History (OH-SMArt[4]) and conversations in the medical domain (Tejedor García et al., 2022).

Benchmark data: As a reference or baseline dataset, we choose the N-Best 2008 evaluation corpus (Van Leeuwen et al., 2009). N-Best contains broadcast news speech and telephone conversations, in Netherlands-Dutch and Flemish, and is reasonably representative for the data researchers encounter in an audiovisual media archive. This dataset was also used for the evaluation of the ASR system that was initially provided a number of years ago in the CLARIAH research infrastructures (Ordelman & van Hessen, 2018). We also evaluated the JASMIN-CGN corpus, an extension of the Spoken Dutch Corpus (CGN) that contains speech from native Dutch/Belgian children, the elderly, as well as non-native speakers. In this paper, the results of ASR on native elderly and non-native adult speech are reported, which occur often in Oral History. As for the speech in the medical domain, 3 datasets have been used: Medicijnjournaal (MJ) (Tejedor García & van der Molen, 2022), Medical Video (MV) material, and sensitive patient-provider conversations. For all datasets, the average performance is reported.

Data Preparation: Standard normalization procedures have been applied such as converting numbers into words, removing punctuation, and removing case distinctions. All normalization steps were stored on the benchmark website as a reference for collaborators and to support fair comparisons between datasets. As for the audio data, they have been either segmented according to the timestamps present in the reference files (in the case of N-Best) or silenced in the regions where no speech is present (in the case of JASMIN-CGN). As for the medical domain, MJ data has been manually annotated according to a protocol detailed in Tejedor García et al. (2022).

ASR Systems and Configurations: The Kaldi_NL system that has been used most frequently in Dutch RIs, is regarded as our baseline system. Kaldi_NL is a collection of DNN-HMM ASR models with speaker diarization developed using the Kaldi toolkit (Povey et al., 2011). Due to the rapid advancements in the field of ASR, it is important to compare its performance with newer systems to see if and for which types of data it should be replaced or updated. For the medical domain data, a fine-tuned Kaldi_NL was

[3] https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/

[4] https://www.uva.nl/en/discipline/conservation-and-restoration/research/research-projects/oh-smart/oh-smart.html

made available, and trained on in-domain data to improve performance.

Given its current popularity and the frequent questions from researchers we receive about its performance, OpenAI's Whisper (Radford et al., 2022) was considered an important ASR system to evaluate. It is a multilingual ASR model that has become popular over the last year due to its high performance across several languages, without additional model training (fine-tuning) needed. The results show the performance of the 'large-v2' and 'large-v3' pre-trained models. Each one is either combined with Voice Activity Detection (VAD) or not. For the medical domain data, only Whisper 'large-v2' without VAD has been evaluated at the moment.

Also, we report on the Massively Multilingual Speech (MMS) engine from Meta AI (Pratap et al., 2023), the most recent development that uses wav2vec 2.0 as the underlying architecture. The version evaluated has been trained on more than 1000 languages. The medical domain data is evaluated instead using a version of wav2vec 2.0 fine-tuned on Dutch (Grosman, 2021).

Metrics: In this paper, only the commonly used Word Error Rate (WER) metric is reported, calculated as the sum of error words output by the model divided by the number of words in the reference text. The lower the metric is, the better the performance. Evaluation time has also been measured for N-Best and JASMIN-CGN, which can be found on the official benchmark website.

3 Results & Discussion

Model	**The Netherlands**		**Flemish**	
	Broadcast News	**Conversational**	**Broadcast News**	**Conversational**
Kaldi_NL	12.6%	38.6%	21.2%	59.4%
Whisper large-v2	10.6%	24.1%	**13.0%**	38.5%
Whisper large-v3	12.5%	25.5%	14.9%	38.4%
Whisper large-v2 + VAD	**10.0%**	**23.9%**	13.6%	37.9%
Whisper large-v3 + VAD	12.3%	25.1%	14.6%	**36.9%**
MMS - 1162 languages	18.5%	42.7%	19.4%	57.7%

Model	**The Netherlands**				**Flemish**			
	Read		**Conversational**		**Read**		**Conversational**	
	N-Nat	**E**	**N-Nat**	**E**	**N-Nat**	**E**	**N-Nat**	**E**
Kaldi_NL	45.3%	20.9%	60.0%	44.0%	43.3%	24.7%	64.4%	47.4%
Whisper large-v2	30.6%	13.7%	77.7%	39.9%	21.0%	16.7%	67.3%	45.4%
Whisper large-v3	62.6%	27.6%	84.5%	51.4%	41.1%	38.7%	79.9%	68.3%
Whisper large-v2 + VAD	**30.0%**	**12.8%**	**51.4%**	**26.8%**	**20.5%**	**14.4%**	**49.3%**	**30.6%**
Whisper large-v3 + VAD	49.4%	25.2%	58.2%	33.6%	50.7%	33.6%	57.9%	44.6%
MMS - 1162 languages	54.0%	28.3%	83.3%	59.9%	35.8%	22.3%	76.7%	60.8%

Model	**Medicijnjournaal**	**Medical Videos**	**pat-prov_test**	**pat-prov_train**
Kaldi_NL	16.1%	28.4%	71.2%	68.5%
Kaldi_NL fine-tuned	-	-	68.0%	-
Whisper large-v2	-	**10.9%**	**57.1%**	**34.1%**
wav2vec2	**12.8%**	24.2%	-	-

Table 1: WER results on N-Best dataset (top table), JASMIN-CGN dataset (middle table), and on medical domain (bottom table). **N-Nat**=Non-native; **E**=Elderly; **pat-prov**=patient-provider conversations.

As the benchmark is a collaborative initiative, the performance versus dataset matrix is *sparse*: not all ASR engines/configurations could be tested on all available data sets. The results can be found in table 1.

Overall, for the Oral History domain, Whisper demonstrates robustness on various styles of speech, different categories of speakers, and on both Flemish and Netherlands Dutch. In contrast, MMS performs the worst overall, indicating a lack in balancing Dutch training material. For the medical domain, both end-to-end models (Whisper and wav2vec2) outperform Kaldi_NL, emphasizing that end-to-end models manage to improve upon the previous state-of-the-art for the medical domain annotation task.

4 Conclusion

Benchmarking ASR engines, configurations and models, and providing evaluation results online such as we are currently doing on GitHub, helps scholars, infrastructure providers and speech researchers in making informed choices about which ASR to use, to facilitate or to improve upon. In order to maximize its potential, first of all, we aim to expand collaborations in the field in terms of providing annotated data sets for speech types that are not evaluated yet, and optionally, running evaluations with new or alternative ASR configurations or models. Especially the provisioning of (small amounts of) annotated data in combination with a semi-automatic evaluation procedure, could be an approach for researchers to obtain performance measures for their data set relatively quickly. Secondly, together with RI providers we will investigate methodologies to incorporate benchmark results into tool registers in a structural, replicable and transparent manner. Finally, together with speech researchers we will investigate how we could optimize the coverage of evaluated speech types and conditions in our benchmark, and for which (typically less common) speech it is required to improve on results provided by current systems and models.

Through this initiative, we encourage researchers and developers of Dutch and Flemish ASR to collaborate by contributing to the Dutch Open Speech Recognition Benchmark with their results. Additionally, we would welcome collaborations with similar initiatives for other languages on a more European level, within the CLARIN context, as well as beyond.

Acknowledgments

This work has been done as part of the CLARIAH-PLUS project funded by NWO (grant 184.034.023), as well as part of *PDI-SSH* projects ”Oral History – Stories at the Museum around Artworks (*OH-SMArt*)” and ”Homo Medicinalis. Recognising the voices of patients to retrace the opinions about medicines (*HoMed*)”.

References

Gref, M., Matthiesen, N., Schmidt, C., Behnke, S., & Köhler, J. (2022). Human and automatic speech recognition performance on german oral history interviews. *arXiv preprint arXiv:2201.06841*.

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in Dutch.

Koolen, M., van Gorp, J., & van Ossenbruggen, J. (2018). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, *34*(2), 368–385. https://doi.org/10.1093/llc/fqy048

Ordelman, R., Melgar, L., Van Gorp, J., Noordegraaf, J., et al. (2018). Media suite: Unlocking audio-visual archives for mixed media scholarly research. *Selected papers from the CLARIN Annual Conference*, *159*, 133–143.

Ordelman, R., & van Hessen, A. J. (2018). Speech recognition and scholarly research: Usability and sustainability. *CLARIN 2018 Annual Conference*, 163–168.

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 Catalog No.: CFP11SRW-USB]. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ..., & Auli, M. (2023). Scaling Speech Technology to 1,000+ Languages.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

Tejedor García, C., & van der Molen, B. (2022). *Homed Transcriptions Medicijnjournaal* (Version 1). Radboud University. https://doi.org/10.34973/dpjc-0v85

Tejedor García, C., van der Molen, B., van den Heuvel, H., van Hessen, A., & Pieters, T. (2022). Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1032–1039.

Van Leeuwen, D., Kessens, J., Sanders, E., & van den Heuvel, H. (2009). Results of the N-Best 2008 Dutch Speech Recognition Evaluation, 2571–2574. https://doi.org/10.21437/Interspeech.2009-677

Automatic analysis of covert hate speech: A case study with a focus on sentiment analysis

Anna Szczepaniak-Kozak
Institute of Applied Linguistics
Adam Mickiewicz University
Poznań, Poland
anna.szczepaniak-kozak@amu.edu.pl

Magdalena Jaszczyk-Grzyb
Institute of Applied Linguistics
Adam Mickiewicz University
Poznań, Poland
magdalena.jaszczyk@amu.edu.pl

Abstract

Our presentation focuses on the findings of a study on covert hate speech detection using sentiment analysis. The study investigates below-the-line (BTL) comments on Ukrainians residing in Poland expressed in Polish. To perform analyses, we used two software packages, *MultiEmo* and *HateSpeech* (developed by CLARIN-PL). Our findings show that on average 30% of the comments feature negative sentiment but the software packets turn out slightly different percentage levels. We also offer recommendations on fine-tuning the software and increasing manual annotation in follow-up studies.

1 Introduction

Covert hate speech relies on using disguised ways to express racism, sexism, homophobia or any bias against a specific community. It can be conveyed via diverse tropes, including metaphors, analogies, proverbs, understatements, sarcastic remarks, as well as dog whistling strategies (Baider 2019; Baider and Constantinou 2020; Bhat and Klein 2020; Åkerlund 2021). In these cases, code words are used instead of racist terms or to imply racial stereotypes as multi-vocal appeals that have distinct meanings to different audiences (Albertson 2015).

Despite attempts to regulate cyberhate through international instruments and social media platforms' own bylaws (Fortuna and Nunes 2018), as well as significant progress in automatic detection of overt hate speech (Waseem 2016; Castaño-Pulgarín et al. 2021), the process has been hindered by the fact that covert hate speech is much more difficult to detect and track. To address this, our study focuses on sentiment analysis, which identifies the positive, negative or neutral stance that seems to characterise online posts. We are testing this method with a focus on online BTL (below-the-line) comments on Polish news articles and blogs about Ukrainians, using two software packages. The aim of our study is to analyse the sentiment of general Internet content in Polish on Ukrainians, to understand the mechanisms of covert hate speech, and enable a broader understanding of such linguistic processes through corpus linguistics approach. Our research question is:
RQ: What is the sentiment of below-the-line (BTL) comments in Polish online discourse on Ukrainians?

2 Methodology

2.1 Software packets used

Current automatic tools that focus on keyword detection, for example, seem not to be sufficient to elucidate covert hateful content (Waseem, 2016: 141). This is why in this study sentiment analysis (SA) was used. SA is defined by Liu (2012: 1) as "the mining of opinions of individuals, their appraisals, and their feelings in the direction of certain objects, facts and their attributes". It is understood as a semi-automatic data analysis aiming at uncovering emotions, including text subjectivity and sarcasm detection (Cambria et al. 2017: 4–5), conveyed by language-related choices made by people producing written or spoken utterances (Medhat et al. 2014: 1093). In our study, we use SA tools *MultiEmo* and *HateSpeech* that have been recently developed by CLARIN-PL (Kocoń et al. 2021a,b,c).

MultiEmo is an open-source software (https://ws.clarin-pl.eu/multiemo) which can recognize sentiment for more than 100 languages. It assigns sentiment at the level of the whole document, individual paragraphs or sentences with a value on the continuum from positive, neutral, ambivalent to negative (Kocoń et al. 2021a).

HateSpeech is also an open-source tool (https://ws.clarin-pl.eu/hatespeech). It detects text offensiveness and ascribes sentiment on the continuum between 0 and 1. The software is prepared in two variants. The first variant uses a generic model for recognizing offensiveness and was trained on averaged text annotation values. The second variant uses a personalized model and for prediction it uses information in the form of a questionnaire filled out by a human against whom it is supposed to predict offensiveness (Kocoń et al. 2021b,c).

2.2 Corpora and method of analysis

In this study, we used corpus data of two types: focus/tailored and reference corpus. The main dataset (henceforth focus corpus) comprises more than 12 million words in total. It compiles anonymised data in Polish extracted from various domains in which there were references to Ukrainians. This main corpus was divided into eight subcorpora depending on the relevant key word they feature. We also used *Polish Web 2019*, a corpus representative of contemporary Polish internet communication, serving as a reference corpus. Its size is 4.2 billion words. The source of these texts is the Polish domain ".pl".

After collecting data, we proceeded to establish the sentiment of the analysed focus corpus, against the reference corpus, through Analysis of the saturation of BTL comments with negative and offensive content by means of MultiEmo and HateSpeech.

3. Analysis of the saturation of particular BTL comments with hateful content

To investigate the sentiment of the collected data and to answer RQ, the focus datasets were fed into *MultiEmo* and *HateSpeech*. In MultiEmo, each comment was labelled with the level of saturation within four sentiment classes as per the figures below: (1) minus: negative; (2) plus: positive; (3) zero: neutral; (4) amb: ambivalent (meaning that a particular text includes both positive and negative aspects that are balanced in terms of relevance). In HateSpeech, each comment is labelled with the level of saturation within two sentiment classes: offensive and inoffensive. To illustrate the analyses enabled by the software, we discuss Example 1, taken from one of our focus subcorpora.

Example 1:
Text in Polish: puknij się pajacu w czółko..........dwa dni i po banderowcach....dorwały się świnie do koryta... (Translation into English: knock yourself on your little forehead, your moron..........two days and away with banderowiecs[1]....the pigs got to the trough...)

This individual entry was labelled as negative in sentiment by MultiEmo (75% of the entry's content is negatively saturated). It is also recognised as offensive by HateSpeech (95% of its content is offensive).

These software packages can also provide us with indications about the sentiment categorisation for entire datasets. The distribution of the sentiment in an exemplary subcorpus ("Banderowiec" subcorpus) is presented in Figure 1. It indicates that around 25% of comments were classified as offensive to a different extent. This example is different from the one above, as it does not indicate the saturation with hateful content only, but rather the scale of negative sentiment of any type.

[1] Banderowiec is a negative ethnonym for Ukraininas in Polish; the word comes from a Ukrainian militia/paramilitary group active during WWII;

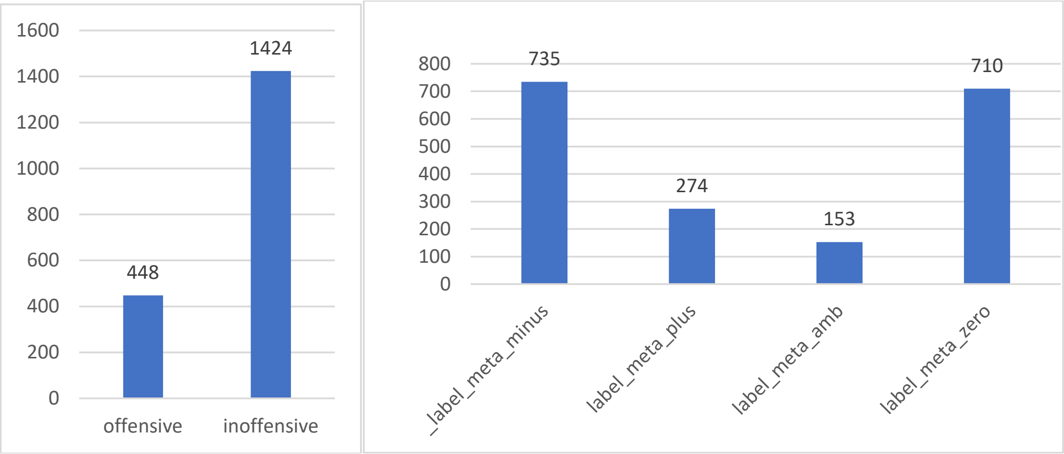


Figure 1: The distribution of sentiment as detected by HateSpeech (left visualisation) and MultiEmo (right visualisation) in subcorpus “Banderowiec”.

In our presentation, we also intend to present the distribution of sentiment in the majority of our focus datasets to answer our research question. The average score for negative sentiment is at 39.96% (MultiEmo), while the scale of offensiveness in the entire data is 23.78% (HateSpeech). This means that offensiveness appears at a lower rate across the subcorpora. This is expected, as these software will have different degrees of sensitivity regarding offensiveness, and negative comments may not necessarily be flagged as inoffensive (more about it in the section below).

4 Conclusions, limitations and recommendations for further research

The software packages we used for SA are very sensitive to the emotional saturation of texts and generated findings which enabled us answering the research question.

Even though the SA tools MultiEmo and HateSpeech provided a very clear picture of the saturation of the dataset with sentiment, including negativity and offensiveness, they both require fine-tuning. In manual annotation checks of our focus corpora, we found that some comments were falsely indicated as inoffensive or offensive (see Example 2 and Tables 1 and 2). This is partly because the software packages do not differentiate between different contexts or metaphorical uses of particular words, as we saw above. For this reason, we also need better annotation guidelines, particularly for detecting “specific subsets of abusive language” (Waseem 2016: 141).

Example 2

Comment in Polish: *Trzy wyjątkowo parszywe świnie dobrane w korcu maku- ukrainiec - banderowiec, rzydówka i folksdojcz*
Translation into English: Three exceptionally mangy pigs, like two peas in a pod – a Ukrainian, bandera, Jewish woman and Volksdeutsch)

MultiEmo turnout for this very negative comment was evaluated as negative at 42.9% and positive in 28.8% (see Table 1).

Table 1. MultiEmo scores for Example 2.

multiemo___label__meta_**minus**_m	multiemo___label__meta_**plus**_m	multiemo___label__meta_**zero**	multiemo___label__meta_**amb**
0.429	0.288	0.206	0.078

HateSpeech turnout was also false positive (74.7%) and the software classified this comment as inoffensive (see Table 2).

Table 2. HateSpeech scores for Example 2.

hatespeech_inoffensive	hatespeech_offensive	is_offensive
0.7470712714	0.2529287286	Inoffensive

Further research into covert online hate speech is necessary because we still know too little about the strategies and linguistic means used to express it. It would be worthwhile to look for keywords which denote a contemptuous attitude, especially because contempt can lead to hatred (Baider and Constantinou 2017; Miceli and Castelfranchi 2018). Contempt is often expressed in sarcastic, ridiculing or mocking messages. Szczepaniak-Kozak's (2023) corpus linguistics study investigates contempt in discourse about Ukrainians living in Poland, however further investigation could help making its detection easier and faster to raise public awareness and enable its prosecution by law enforcement bodies.

Acknowledgements: The authors would like to thank CLARIN-PL, especially prof. Maciej Piasecki, dr. Jan Wieczorek and dr. Tomasz Bernaś for their support. This presentation partially bases on: F. Baider, A. Szczepaniak-Kozak, K. Strani & M. Jaszczyk-Grzyb (in review). *Covert hate speech analysed automatically: A case study with a focus on self-victimisation*.
Research funding: Funded by the British Academy (Pump-priming grant No. PPHE210314) and the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, projects no. POIR.04.02.00-00C002/19 and POIR.01.01.01-00-0288/22 (as part of the authors' collaboration with CLARIN-PL).

References

Åkerlund, M. (2021). Dog whistling far-right code words: the case of 'culture enricher' on the Swedish web. *Information. Communication & Society* 25. 1808–1825.

Albertson, B. L. (2015). Dog-whistle politics: Multifocal communication and religious appeals. *Political Behavior* 37(1). 3–26.

Baider, F. & Constantinou, M. (2017). Burn the antifa traitors at the stake…. Transnational political cyber-exchanges, proximisation of emotions. In Istvan Kecskes & Stavros Assimakopoulos (eds.), *Current Issues in Intercultural Pragmatics*, 75–102. Amsterdam: John Benjamins.

Baider, F. & Constantinou, M. (2020). Covert hate speech: A contrastive study of Greek and Greek Cypriot online discussions with an emphasis on irony. *Journal of Aggression Language and Conflict* 8(20). 262–287.

Baider, F. (2019). Le discours de haine dissimulée; le mépris pour humilier. *Déviance et société* 43(1). 71–100.

Baider, F. (2020). Pragmatics lost? Overview, synthesis and proposition in defining online hate speech. *Pragmatics and Society* 11(2). 196–218.

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega L. M. T. & Herrera López M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior* 58, 101608, https://doi.org/10.1016/j.avb.2021.101608 (accessed 10 April 2024).

Fortuna, P. & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51(4). 1–30.

Kocoń, J., Miłkowski, P. & Kanclerz, K. (2021a). Multiemo: Multilingual, multilevel, multidomain sentiment analysis corpus of consumer reviews. *International Conference on Computational Science.* Cham: Springer. https://www.iccs-meeting.org/archive/iccs2021/papers/127430291.pdf (accessed 02 October 2023).

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T. & Kazienko, P. (2021b). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management* 58(5). 102643. https://doi.org/10.1016/j.ipm.2021.102643

Kocoń, J., Gruza, M., Bielaniewicz, J., Grimling, D., Kanclerz, K., Miłkowski, P. & Kazienko, P. (2021c). Learning personal human biases and representations for subjective tasks in Natural Language Processing. *International Conference on Data Mining (ICDM)*. 1168–1173.

Liu, B. (2012). Sentiment analysis: A fascinating problem. In Hirst Graeme (ed.), *Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies*, 1–8. Cham: Springer. https://doi.org/10.1007/978-3-031-02145-9_1

Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5(4). 1093–1113.

Miceli, M. & Castelfranchi, C. (2018). Contempt and disgust: The emotions of disrespect. *Journal for the Theory of Social Behaviour* 48. 205–229.

Szczepaniak-Kozak, A. (2023). Deconstructing hate speech messages by means of dual character concepts: Finding evidence of newly emerging contemptuous meanings with recourse to philosophical concepts and corpus linguistics. In Hadrian Lankiewicz (ed.), *Extending research horizons in applied linguistics: Between interdisciplinarity and methodological diversity*, 58–73. Sheffield: Equinox.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142. Austin, Texas. Association for Computational Linguistics https://doi.org/10.18653/v1/W16-5618

Different Registers, same Learners: Towards a Multi-Register Corpus of Learner English

Sylvie De Cock
Centre for English
Corpus Linguistics
UCLouvain, Belgium
`sylvie.decock@uclouvain.be`

Gaëtanelle Gilquin
Centre for English
Corpus Linguistics
UCLouvain, Belgium
`gaetanelle.gilquin@uclouvain.be`

Sylviane Granger
Centre for English
Corpus Linguistics
UCLouvain, Belgium
`sylviane.granger@uclouvain.be`

Pauline Jadoulle
Centre for English
Corpus Linguistics
UCLouvain, Belgium
`pauline.jadoulle@uclouvain.be`

Magali Paquot
Centre for English
Corpus Linguistics
UCLouvain, Belgium
`magali.paquot@uclouvain.be`

Abstract

This paper presents a learner corpus compilation project carried out within the frame of the CLARIN Knowledge Centre for Learner Corpora. The corpus includes data representing different registers (both written and spoken) produced by the same group of learners. This specificity makes it possible to control for individual styles when investigating the effect of register on the linguistic features of learner language, unlike most register studies in learner corpus research which rely on different registers produced by different groups of learners. When completed, the learner corpus will be made available through the Corpor@UCLouvain platform, which is recorded in the CLARIN Virtual Language Observatory.

1 Introduction

Register variation is a crucial aspect of language production. Depending on the context in which it is used and the communicative purposes that it serves, language tends to display distinctive characteristics, which may have to do with lexis, but also phraseology or syntax, among others (see, e.g., Biber 2012). Corpora, in particular, have raised awareness of register variation and have provided insights into linguistic patterns associated with different registers. Specific methods based on corpora have also been developed to study register variation, most notably multi-dimensional analysis (Biber 1988). This paper focuses on the corpus study of register variation in learner language. It considers the limitations of register studies in learner corpus research and describes the design of a learner corpus meant to represent multiple registers produced by the same learners and compiled within the frame of the CLARIN Knowledge Centre for Learner Corpora (CKL2CORPORA).

2 Register Studies in Learner Corpus Research

While register is important for any type of language, it is particularly relevant to learner language, because learners may not show the same register awareness as native or expert writers/speakers (cf. Gilquin & Paquot 2008). Learner corpus research has taken register into account in the sense that studies are carried out on the basis of learner corpora representing certain registers (e.g. telecollaborative discourse in Vyatkina 2012). However, studies comparing learner language registers are still relatively rare. One reason for this is that, until recently, learner corpora represented only a small range of registers, most notably argumentative essays for writing (as in the International Corpus of Learner English; Granger et al. 2020) and interviews for speech (as in the Louvain International Database of Spoken English Interlanguage; Gilquin, De Cock, & Granger 2010). Over the last decade or so, however, learner corpora representing various registers have started to become available, including language for specific purposes learner corpora (e.g. the Active Learning of English for Science Students Learner Corpus; Allen 2012) and learner translation corpora (e.g. the Multilingual Student Translation corpus; Granger & Lefer 2020). Some learner corpora have also appeared that combine several text types, although the range tends to be rather limited (e.g. only written texts in the Tracking Written Learner Language corpus; Hasund, Drange, & Torjusen 2022).[1]

Using such learner corpora, a few studies have compared two or three registers, such as speech and writing in Fuchs, Götz, & Werner (2016) or argumentative essays and research papers in Larsson, Paquot, & Biber (2021). A couple of studies have also sought to situate a learner language register among various native/expert registers by comparing their linguistic features (cf. Aguado-Jiménez, Pérez-Paredes, & Sánchez 2012). Importantly, when learner corpus researchers have compared different registers, it has mainly been on the basis of texts produced by different learners (e.g. argumentative essays produced by one group of students and interviews produced by another group). A possible issue with this method is that individual writers'/speakers' styles may affect the comparison of registers. In other words, differences between registers may appear simply because different writers/speakers use language differently.

3 A Multi-Register Learner Corpus

As part of a project on register variation and learner language, we have been compiling a learner corpus made up of texts from different registers produced by the same learners. This corpus will make it possible to compare registers while controlling for individual styles, and to investigate the effect of register on the linguistic features of learner language.

The learner data are collected at UCLouvain among (mainly) French-speaking learners of English who are students in their second year of English major studies. These students are required to produce written and spoken texts in English representing different registers. The written registers include (i) a career readiness essay in which students discuss the career they have chosen as well as its potential benefits and drawbacks, (ii) a cover letter which students write in response to a (student) job advertisement, and (iii) a persuasive essay which requires students to propose a solution to a problem they have chosen and persuade readers to agree with their solution. The spoken registers include (i) a non-interactive, monologic task in which students have to explain where they see themselves in 5 years' time, professionally speaking, (ii) an interactive task which takes the form of a job interview with the teacher, and (iii) an interactive task in pairs of students which involves debating the topics they chose for their persuasive essays. It will be noticed that the written and spoken registers are closely linked to each other: the career readiness essay is related to the monologue on students' professional future, the cover letter is the written equivalent of the job interview, and the persuasive essay corresponds to the debate in pairs of students. This should enhance the comparability of the written and spoken registers.

We have also made special efforts to collect rich metadata about the learners and the tasks. Our starting point has been Frey et al.'s (2023) and Paquot et al.'s (2024) Core Metadata Schema for Learner Corpora, whose aim is to increase the standardization of learner corpus metadata. We have included information about, among others, learners' knowledge of languages, their exposure to English in different situations, but also their literacy. Detailed information about the tasks (e.g. instructions, time constraints and use of language reference tools for writing) and the registers (e.g. communicative purposes, settings, number of addressees) is also provided. A learner ID makes it possible to link together all texts produced by one and the same student.

[1] See CLARIN's L2 Learner Corpora resource (https://www.clarin.eu/resource-families/L2-corpora) or the Centre for English Corpus Linguistics' Learner Corpora around the Word webpage (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html) for a list of existing learner corpora.

So far, data have been collected among one cohort of students, representing about 60 learners. The data collection will continue among future cohorts of students, until a sufficient amount of data has been collected. Once completed, we aim to make the learner corpus available to the research community, using a FAIR (Findable, Accessible, Interoperable, and Reusable) approach (see Lindström Tiedemann, Lenardič, & Fišer 2018 and König, Frey, & Stemle 2021). The data will be released in open access format through the Corpor@UCLouvain platform, which is recorded in the CLARIN Virtual Language Observatory (VLO).

This learner corpus is one of the first corpus compilation projects at UCLouvain since CKL2CORPORA was officially recognized in November 2022. The project brings together several members of CKL2CORPORA. It thus gives us the opportunity to carry out some of CKL2CORPORA's missions, such as testing and improving Paquot et al.'s (2024) Core Metadata Schema for Learner Corpora and establishing best practices for data formats (including speech transcriptions).

4 Conclusion

The compilation of our multi-register corpus of learner English, carried out within the frame of CKL2CORPORA, is meant to provide researchers with the necessary resources to study how learner language varies according to register. Not only does it represent an improvement on the resources that are usually exploited for the investigation of registers in learner language, including different registers produced by the same group of learners rather than by different groups of learners, but it also contributes to the establishment of best practices in aspects such as metadata standardization and data formats.

Acknowledgements

We acknowledge the financial support of UCLouvain's Hoover Seed Fund for this project.

References

Aguado-Jiménez, P., Pérez-Paredes, P., & Sánchez, P. (2012). Exploring the use of multidimensional analysis of learner language to promote register awareness. *System*, *40*(1), 90-103.

Allen, D. (2012). Active Learning of English for Science Students (ALESS): A personal introduction. *Proceedings of New Approaches to English Language Education for Students of Science and Engineering in Japan, June 10, 2010*. Published March 26, 2012, pp. 33-37. Waseda University.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*(1), 9-37.

Frey, J.-C., König, A., Stemle, E. W., & Paquot, M. (2023). Core Metadata Schema for L2 data. Paper presented at the 2nd Conference of the European Second Language Association (EUROSLA), University of Birmingham, 30 August 2023 – 02 September 2023.

Fuchs, R., Götz, S., & Werner, V. (2016). The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing. In Werner, V., Seoane, E., & Suárez-Gómez , C. (eds) *Re-assessing the Present Perfect*, pp. 297-338. De Gruyter.

Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Presses universitaires de Louvain.

Gilquin, G. & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction 1*(1), 41-61.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.

Granger, S. & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: A resource for translation teaching and research. *Language Resources and Evaluation*, 54, 1183-1199.

Hasund, I. K., Drange, E.-M., & Torjusen, H. M. H. (2022). Background, context and studies of the TRAWL corpus. *Nordic Journal of Language Teaching and Learning*, *10*(2), I-XVIII.

König, A., Frey, J.-C., & Stemle, E. W. (2021). Exploring reusability and reproducibility for a research infrastructure for L1 and L2 learner corpora. *Information*, *12*, 199.

Larsson, T., Paquot, M., & Biber, D. (2021). On the importance of register in learner writing: A multi-dimensional approach. In E. Seoane & D. Biber (eds) *Corpus-based approaches to register variation*, pp. 235-258. John Benjamins.

Lindström Tiedemann, T., Lenardič, J., & Fišer, D. (2018). L2 learner corpus survey : Towards improved verifiability, reproducibility and inspiration in learner corpus research. Proceedings of *CLARIN Annual Conference 2018, Pisa, Italy, 8-10 October 2018*, pp. 146-150.

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024). The Core Metadata Schema for Learner Corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research*, *10*(2).

Vyatkina, N. (2012). Applying the methodology of learner corpus analysis to telecollaborative discourse. In Dooly, M. & O'Dowd, R. (eds) *Researching online foreign language interaction and exchange: Theories, methods and challenges*, pp. 267-303. Peter Lang.

CLARIN in the Italian Open Science Cloud: Landscaping and Community Engagement

Roberta Bianca Luzietti
CNR-ILC / University of Pisa
robertabianca.luzietti@cnr.it

Valeria Quochi
CNR-ILC
valeria.quochi@cnr.it

Roberta Ottaviani
CNR-ILC
roberta.ottaviani@cnr.it

Daniele Carpita
ILC-CNR
daniele.carpita@cnr.it

Riccardo Del Gratta
CNR-ILC
riccardo.delgratta@cnr.it

Monica Monachini
ILC-CNR
monica.monachini@cnr.it

Abstract

This contribution is part of the H2IOSC project, supported by the Italian PNRR European Fund, in which the Italian CLARIN node collaborates with DARIAH, E-RIHS, and OPERAS to build an Italian Open Science Cloud. The paper presents an overview of two key project activities aimimg at landscaping the Italian resource panorama and increasing the Italian research community's involvement. On the one hand, CLARIN-IT has benefited from CLARIN ERIC central services such as Virtual Language Observatory and Resource Families to gather information on the type and status of resources available that might be of interest of the Italian research community. On the other hand, through the H2IOSC activities CLARIN-IT is working to increase and strengthen the influence and use of CLARIN services within the Italian linguistics community.

1 Introduction

This paper presents the main goals and initial outcomes of the "landscaping and building communities" activity[1] conducted by CLARIN-IT in collaboration with DARIAH, E-RIHS, and OPERAS within the Humanities and cultural Heritage Italian Open Science Cloud (H2IOSC) project. Funded by the Italian PNRR European Fund, H2IOSC aims to establish an Italian Open Science Cloud based on a community-sourced approach and create a virtual environment for sharing and accessing resources and services across different scientific disciplines.

A federation in which RIs collaborate to provide greater availability and interoperability of resources is not a novelty (Broeder et al., 2020). There are, in fact, successful experiences such as the SSHOC[2] project at the international level, and similar initiatives at a national level, such as Text+[3], the consortium of the National Research Data Infrastructure (NFDI) in Germany. H2IOSC in Italy will maximize the collaborative efforts among four Italian Social Science, Humanities, and Cultural Heritage Science infrastructures to increase their influence among research communities and leverage the possibilities of digitization in research, teaching, and transfer, establishing a common data culture for the sustainability of H2IOSC for the next ten years. At the heart of the H2IOSC workflow there are i) a comprehensive investigation of the desired resources and priorities regarding the communities of interest and ii) a review of the resources already available from the partnering research infrastructures repositories, such as ILC4CLARIN, and online repositories and catalogs.

Significant efforts are dedicated to the "landscaping and building communities" activity, led by the Italian CLARIN group at CNR-ILC. In this context "landscaping" refers to the identification and mapping of tools, services, training materials, datasets, publications, and workflows—whether commonly used or newly created within the Italian context (Del Gratta et al., 2021). The outcomes of this landscaping effort will be made accessible through the Research Infrastructures' (RI) repositories via an H2IOSC Observatory linked to the H2IOSC Marketplace discovery portal. The implementation of the observatory will serve to monitor: (i) the status and developments of resources (to be) included in the H2IOSC

[1] https://www.h2iosc.cnr.it/the-project/(last accessed 22/04/2024)

[2] https://marketplace.sshopencloud.eu (last accessed 22/04/2024)

[3] https://www.nfdi.de/textplus/?lang=en (last accessed 22/04/2024)

Marketplace; (ii) the research community needs and priorities; and (iii) the panorama (of language resources and technologies) in Italy over time. While information gathering activities from the community are a fundamental aspect of all infrastructural projects, the distinctive contribution of H2IOSC lies in its robust methodological foundation, employing a mixed methods approach and encouraging the direct involvement and participation of research communities (Tashakkori and Creswell, 2007).

The paper is structured as follows: section 2 details the methods adopted in the landscaping focusing on the questionnaire(s) and focus group, section 3 presents preliminary considerations based on the initial surveying activity results, finally in section 4 we present our conclusions and future developments.

2 Landscaping and Building Communities Methodology

Community engagement and information gathering are the key elements leading the four RIs representatives working in the "landscaping" research group for the implementation of a comprehensive survey aiming to target existing projects, resources, tools, communities, best practices, and standards in the Italian research panorama[4]. A unified landscaping strategy was developed to optimize both information gathering and community engagement, guided by the following goals:

i) understanding the degree of knowledge and interest of research communities to participate in infrastructural activities,
ii) acquiring information on the type of resources, tools, and services identified, as well as their level of FAIRness,
iii) identifying the most used and innovative resources, tools, and services for prioritization of availability through repositories and showcase in the Marketplace, and
iv) aligning the offerings of each infrastructure with the interests and requirements of research communities.

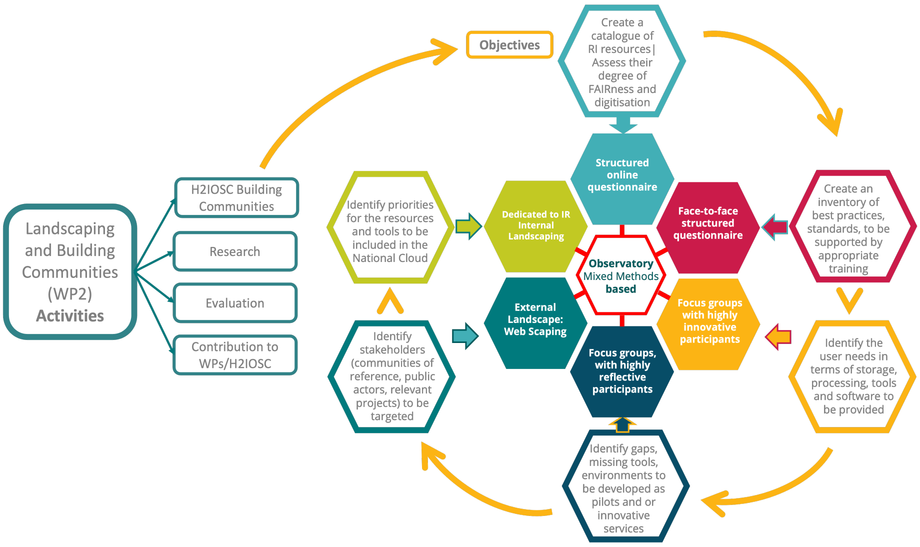

Objectives
Create a catalogue of RI resources| Assess their degree of FAIRness and digitisation
Landscaping and Building Communities (WP2) Activities
H2IOSC Building Communities
Research
Evaluation
Contribution to WPs/H2IOSC
Identify priorities for the resources and tools to be included in the National Cloud
Structured online questionnaire
Dedicated to IR Internal Landscaping
Face-to-face structured questionnaire
Create an inventory of best practices, standards, to be supported by appropriate training
Observatory Mixed Methods based
External Landscape: Web Scaping
Focus groups with highly innovative participants
Focus groups, with highly reflective participants
Identify stakeholders (communities of reference, public actors, relevant projects) to be targeted
Identify the user needs in terms of storage, processing, tools and software to be provided
Identify gaps, missing tools, environments to be developed as pilots and or innovative services

Figure 1: H2IOSC Mixed Methods Landscaping Methodology (Luzietti et al., 2024: 2)

To achieve this, a mixed-methods approach was adopted to gather both quantitative and qualitative data, further enriched through an active participatory research strategy (Luzietti et al., 2024). Fig. 1 illustrates the methodology and workflow of the H2IOSC landscaping activities, which fundamentally comprise

[4] https://landscape2024.esfri.eu/

three instruments: survey questionnaires, focus groups, and data collection from web sources[5], each designed for distinct but intersecting objectives.

2.1 Questionnaire and Focus Groups

The first instrument implemented during the project's first year was an online questionnaire (Krosnick, 2018; Nardi, 2018). This task was particularly complex due to the heterogeneity of the targeted research communities and the simultaneous need to stimulate their interest and engagement to ensure effectiveness. To gather feedback and ensure the questionnaire's clarity and relevance, we first submitted a test version to a group of experts representing the diverse research interests of each infrastructure (e.g., linguists, archaeologists, philologists). Despite these precautions, the initial version of the questionnaire, released in September 2023, did not meet the project's expectations. Early results showed a low participation rate (20 percent), particularly among the Italian Linguistics community. To address this issue, a few months later the Italian CLARIN Consortium organized an in-person event to discuss the challenges and obstacles that might have discouraged participation. The meeting revealed that the most problematic section of the questionnaire involved the description of key resources and technologies used to assess their FAIRness level (e.g., resource type, location, presence of PID/DOI/handle, format, license, and associated publications). Respondents found this section difficult to complete, which contributed to the low participation. To resolve these issues, we opted for splitting the survey into two separate questionnaires. The first focused on collecting general information on the respondents, such as career level, knowledge of IRs, attitudes and interest in IRs and H2IOSC activities, and only two broad questions about the resources and tools used. The second questionnaire was, instead, dedicated to a more detailed but simplified description of the aforementioned used and created resources and technologies. To encourage participation without overwhelming respondents, we allowed them to complete the second questionnaire online, in a face-to-face interview, or to opt out entirely. After applying the necessary revisions, we released a second version of the questionnaire in July 2024 [6].

The second strategy involved the conduct of focus groups (Bezzi, 2020; Frisina, 2010) a research technique that provides access to meanings, processes, and applications of socially shared norms (norms that influence the judgments, behaviors, and actions of individuals in research environments) thoughts and ideas that the individual, more often than not, has no way or opportunity to bring out. In the project's second year, we scheduled two types of focus groups each composed of students, early-career researchers, experienced professionals, professors and senior researchers. In the first "visionary" focus group the discussion is aimed to focus on future-oriented topics, particularly the participants' expectations regarding the availability of new or restored resources, tools, and training initiatives. In contrast, the second "experienced" group is aimed for a more balanced discussion, combining both forward-looking ideas and traditional considerations, with an emphasis on the long-term sustainability of resources, project services, and workflows through the H2IOSC Marketplace. By the end of July 2024, took place the first focus group with participants selected by the four RIs, who were not directly involved in H2IOSC[7]. The methodology employed two moderators: one to present the discussion prompts and direct the conversation, and the other to supervise and guide the discussion when necessary. After a brief presentation of the objectives and introductions, the moderators initiated a series of prompts to assess participants' knowledge and past use of RIs, as well as their interest in becoming more involved in RI activities and the H2IOSC project.

3 Preliminary Findings

Preliminary findings from the two questionnaires and initial focus groups revealed significant variation in community participation across the four research infrastructures. Comparing the first and second ver-

[5] For organizational purposes, we diversified the collection of information from web sources into two categories: "internal landscaping" and "external landscaping". The first one refers to exploring repositories and catalogs developed by the partnering RIs, whereas the second refers to other repositories and catalogs, publications, and conference proceedings).

[6] The questionnaire is available in Italian at https://www.h2iosc.cnr.it/survey-landscaping-and-building-communities/ (last accessed 6/09/2024)

[7] The first focus group was prepared and conducted by Nicola Giampietro and Marta Caradonna from OPERAS-IT.

sion of the questionnaire we also noted a 15 percent increase in responses along with positive feedback from the community. Participants' disciplinary backgrounds were of course different but in line withe RIs (e.g., ERC PE4, PE5, SH4, and SH5) whereas for career levels we noted that the majority of respondents were senior researchers, and professors. Interestingly, most respondents identified as users rather than developers of resources and tools such as databases (linguistic, semantic, relational, analytical), archives, corpora (written and oral), 2D/3D models, annotation software, and tokenizers—many of which are neither openly accessible nor reusable. For this reason, the CLARIN-IT landscaping unit is collaborating closely with the H2IOSC training group for the organization of events aimed at showcasing what the CLARIN ERIC network and repositories can offers. Moreover, these events are aimed to be held not only at general university gatherings but also at smaller research units currently working on projects, with the goal of guiding and assisting researchers along the data acquisition, management, and deposit. On a positive note, the second version of the questionnaire showed increased awareness of the RIs among research communities. More importantly, respondents expressed a strong interest and willingness to participate in training and outreach activities, as well as a growing intention to share their scientific outputs.

Also the results from the focus groups revealed several key insights. Most participants were already familiar with at least one research infrastructure and reported actively using its services and only one participant declared to have experience with the development of new tools and resources. All participants indicated to publish using Open Access systems and expressed a strong willingness to become more FAIR compliant in the future despite the possibility of encountering some challenges. Importantly, there was a shared interest in receiving training on how to effectively use digital research infrastructures and to learn more about the upcoming services that the H2IOSC project will offer. The most important outcome of this first focus group was that being encouraged by the open format of the discussion, participants took time to discuss their experience with using digital tools and resources, developing digital tools, and creating digital resources. The conversation focused on these topics and revealed an important insight that methods such as the questionnaires could not capture. Another interesting point was that for some research fields, especially within the humanities, the lines between using and creating digital resources are often blurred. For instance, linguists employing corpora to train large language models (LLMs) affirmed to be not just users of existing resources but also creators of new ones since trained models can become new digital and reusable resources. This close relationship between resource use and production is particularly important in fields like computational linguistics, where researchers are both consumers and producers. Their work exemplifies how the act of using a resource often leads to the creation of something new, thus contributing to the broader digital research ecosystem.

4 Conclusion and Future Developments

This contribution focuses on the role of CLARIN-IT within the H2IOSC project in building communities and providing a panoramic overview of the Italian research landscape. Compared to other research infrastructures, CLARIN-IT was advantaged by its ability to leverage established services such as the Virtual Language Observatory (VLO), Resource Families[8], and conference proceedings. These tools facilitated the identification and mapping of the newest, most widely used, and well-known resources in the Italian research landscape. For community engagement, CLARIN-IT could also rely on its consortium members and representatives to effectively collaborate with key Italian linguistic associations, enhancing outreach and engagement.

The initial results from the questionnaire and focus groups reveal a positive trend in awareness and participation in research infrastructures, particularly among senior researchers. However, we aim to further increase participation, especially from students and early-stage researchers, as their involvement is crucial for the long-term success and sustainability of the H2IOSC initiative. The mixed-methods approach adopted in this project has proven effective in gathering comprehensive feedback from the research community, allowing for the refinement of strategies and tools to better serve their needs. Looking

[8] https://www.clarin.eu/content/virtual-language-observatory-vlo;https://www.clarin.eu/resource-families (last accessed 22/04/2024)

ahead, continuous refinement of community-building strategies and ongoing collaboration with research infrastructures will be essential to ensuring that the H2IOSC Marketplace evolves into a valuable, sustainable resource hub. By aligning infrastructure offerings with the interests and needs of the research community, H2IOSC aims to contribute significantly to Italy's Open Science efforts, fostering greater cross-disciplinary collaboration and innovation. CLARIN-IT's early successes in resource identification and community engagement provide a strong foundation for future efforts.

Acknowledgments

This research was carried out as part of the H2IOSC Project - Humanities and cultural Heritage Open Science Cloud funded by the European Union - NextGenerationEU - NRRP M4C2 - Project code IR0000029 - CUP B63C22000730005.

References

Bezzi, C. (2020). *Fare ricerca con i gruppi. guida all'utilizzo di focus group, brainstorming, delphi e altre tecniche* (Vol. 12). Milano: Franco Angeli.

Broeder, D., Eskevich, M., & Monachini, M. (2020). Proceedings of the workshop about language resources for the ssh cloud. *Proceedings of the Workshop about Language Resources for the SSH Cloud*.

Del Gratta, R., Goggi, S., Pardelli, G., & Calzolari, N. (2021). The lre map: What does it tell us about the last decade of our field? *Language Resources and Evaluation*, *55*, 259–283.

Frisina, A. (2010). *Focus group, una guida pratica.* Bologna: Il mulino.

Krosnick, J. A. (2018). Questionnaire design. *The Palgrave handbook of survey research*, 439–455.

Luzietti, R. B., Spadi, A., Giampietro, N., Giacomo, M., Caravale, A., D'Eredità, A., Caradonna, M., Moscati, P., Quochi, V., Monachini, M., & Degl'Innocenti, E. (2024). Digital humanities and heritage science: Moving from landscaping to a dynamic research observatory in an open science cloud. In A. Di Silvestro & D. Spampinato (Eds.), *Me.te. digitali. mediterraneo in rete tra testi e contesti, proceedings del xiii convegno annuale aiucd*. AIUCD Associazione per l'Informatica Umanistica e la Cultura Digitale.

Nardi, P. M. (2018). *Doing survey research: A guide to quantitative methods*. Routledge.

Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods.

The CLARIAH FAIR Vocabulary Registry

Kerim Meijer
KNAW Humanities Cluster

Menzo Windhouwer
KNAW Humanities Cluster

{kerim.meijer,menzo.windhouwer}@di.huc.knaw.nl

Abstract

Vocabularies in various forms play a crucial role in achieving the goals of the FAIR principles. However, even though some initiatives try to identify and provide access to them, users from a given community may not yet find a clear way to find them, reuse them, or share their own vocabularies. In this abstract we present the CLARIAH FAIR Vocabulary Registry, which is a web service created to serve both the CLARIN and CLARIAH communities. This semantic artifact registry is created to make vocabularies FAIR themselves. In the first sections we introduce the concepts and review previous work in creating vocabulary registries. In section 4 we detail the architecture of the registry we created; in Section 5 we include a self-assessment of this registry and in Section 6, we outline the future work to make this registry fully functional, not only with an RDF focus, but open to include support for artifacts related to (legacy) data technologies.

1 Introduction

Since the publication of the article on FAIR principles [Wilkinson et al., 2016] (European) infrastructures, e.g., EOSC, and communities, e.g. CLARIN and CLARIAH, are busy making them actionable. For some of the FAIR principles, e.g., the **F** of *Findable*, this is relatively straightforward and there are long standing practices in different communities, e.g., handing out persistent identifiers (PIDs) to metadata and/or resources. But the **I** of *Interoperable* turns out to be quite unwieldy still. A key term in this principle is *Vocabularies*. In the next section we focus on this term and its role in FAIR. In section three we review previous efforts and literature on the topic. In section four we introduce the CLARIAH[1] FAIR Vocabulary Registry[2], which is a practical implementation that contributes to achieving the goals of making information about and the vocabularies related to a FAIR resource also FAIR.

2 The Role of Vocabularies in the FAIR Principles

The FAIR principles explicitly recommend that, for metadata to be FAIR, it has to use vocabularies that follow the FAIR principles (principle **I2**). Thus, both metadata and data should (re)use vocabularies that are themselves available in a FAIR manner. But what are vocabularies? They range from simple lists of terms (a pick list, for example), to more complex structures such as ontologies. From a semantic web perspective (which the FAIR principles often take) vocabularies are often represented in the Resource Description Framework format (RDF[3]), describing properties, classes of resources and relationships between them [McBride, 2004]. In this sense, the vocabulary describes the structures made up of interconnected elements (i.e., the graphs) that can be created by ”triples” using a specific RDF vocabulary. These kind of vocabularies help in **I3** ”(meta)data include qualified references to other (meta)data”, i.e., they provide the qualification of the references to other metadata or data. In the questions of the FAIR Implementation Profile[4] these vocabularies are referred to as model(s) and/or schema(s). Moreover, the *Interoperable* principles actually refer to even more vocabularies. For instance in principle **I1** ”(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.” Such a language for knowledge representation is nowadays often expressed using a RDF vocabulary, e.g., OWL[5] or SKOS[6]. Triples form the knowledge assertions made about a resource and the formal reasoning paradigm underlying that specific knowledge representation determines which entailments can be made on basis of these triples and how the consistency of the embodied knowledge can be validated. This leaves us with the following semantic artefacts (all related to the term *Vocabulary*) which could/should be made explicit to make a resource *Interoperable* in a FAIR manner: 1) knowledge representation languages (e.g., OWL), 2) structured vocabularies (e.g., a particular taxonomy in SKOS), 3) models and schemas (e.g., a schema in RDFS). And these vocabularies (also know as semantic artefacts or FAIR enabling resources) should themselves be FAIR.

3 Related Work

Even before the rise of the FAIR principles, there were significant efforts in creating catalogs and shared infrastructure for, what are now called, semantic artefacts. Other vocabulary registry initiatives are for instance: the ”Linked Open Vocabularies (LOV)”[7], ”dbpedia Archivo”[8] (focusing solely on Linked Open Data Vocabularies), the ”RDA Catalogue of metadata standards”[9] and the ”DCC List of metadata standards”[10] (only for metadata schemas and not for controlled vocabularies); the ”Open Metadata Registry”[11], the ”FAIR sharing”[12] terminology artifacts; ”ontoPortal”[13] and ”Bartoc”[14]. Within the CLARIN domain ”ISOcat” [Kemps-Snijders et al., 2008] and its successor the ”CLARIN Concept Registry”[15] are also such registries. Many of these have a strong Linked Data focus. However, it is unlikely that the abundance of data that is available in the linguistic community will all be transformed into RDF. This does not mean that they can never meet the Interoperability principles of FAIR. As their models and schemas are often also available in well established expressive language such as XML Schema[16] or in a SQL Data Definition Language[17]. However, special care needs to be taken then to connect these models and schemas to controlled vocabularies and knowledge representation languages. Fortunately, this conversion can still be done through Linked Data technology. For example, by using CSV on the Web[18]. This is a path CLARIN already took earlier with the stalled development of SCHEMAcat and RELcat [Windhouwer and Schuurman, 2014]. The ”CLARIAH FAIR Vocabulary Registry” that we developed was motivated by this need to specify the FAIR Interoperability (also known as the FAIRification) of both RDF and non-RDF (meta)data.

4 The CLARIAH FAIR Vocabulary Registry

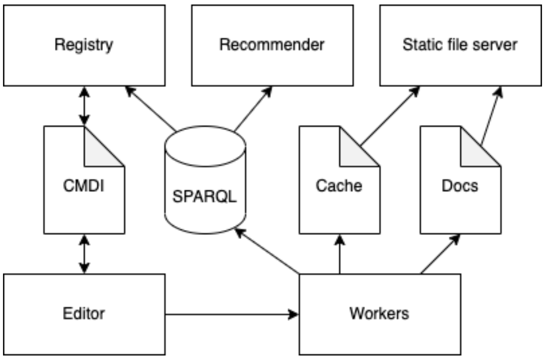


Figure 1: Architecture of the vocabulary registry

The ”CLARIAH FAIR Vocabulary Registry” is a web service where the target users can find FAIR vocabularies to reuse them by selecting schemas and/or controlled vocabularies to model their datasets. This registry will gather relevant vocabularies that can be used by research or cultural heritage projects in the humanities and social sciences carried out by the CLARIAH and CLARIN communities (and collaborating institutions). What distinguishes the CLARIAH FAIR vocabulary registry from some of the initiatives mentioned before is that we aim to: (1) do semi-automatic recommendation work, which include automatic processes in the selection and processing, but also involves experts in the selection and recommendation of the vocabularies, (2) to formalize the characteristics that make a vocabulary FAIR, (3) to serve a clear user group, which gives advantages when selecting vocabularies to include and datasets to link to; also to have a closer relation with users during the development and evaluation of the registry. The architecture of the vocabulary registry is outlined in Figure 1. At its core is a ”Component Metadata Infrastructure” (CMDI)[19] profile defining the required vocabulary metadata and an **editor** facilitating metadata management. Each vocabulary description adheres to this CMDI profile, serving as the input for the vocabulary **workers**. The workers are written in Python using the Celery distributed task queue manager[20]. These workers perform a number of tasks, such as caching vocabulary objects (in a SPARQL[21] store). The **registry** itself is also written in Python as an API service, complemented by a web graphic user interface (GUI) developed in React. Leveraging data from the SPARQL store, the registry empowers users to explore and search through the vocabularies. Cached objects and documentation are accessible via a NGINX[22] **static file server**. Furthermore, the **vocabulary recommender** is a tool written in JavaScript that also utilizes both the stored vocabulary metadata and the vocabularies within the SPARQL store to recommend specific vocabularies based on user search queries.

4.1 Caching

A pivotal aspect of the vocabulary registry is to keep track of the vocabulary history by recording the versions that were made available to the public. Operating as a central hub within CLARIAH, the registry serves as a repository for maintaining these vocabularies. For every version, we record a link in the registry, providing access to the distribution of the specific vocabulary version whenever feasible. To ensure the integrity of the vocabulary history, we proactively store a copy of each distribution in the registry cache. This proactive approach enables users to revert to the cache as an archive in scenarios where the vocabulary is no longer accessible through the recorded link. This includes instances where funding for maintaining the vocabulary expires or when the vocabularies are transferred to another party for maintenance.

4.2 APIs

The registry cache serves as a multi functional resource by also facilitating the enrichment and enhancement of metadata descriptions, but also the creation and provision of *stack*-specific APIs. To delve deeper into understanding a vocabulary, we adopt a systematic approach by categorizing it into distinct *’stacks’*. This categorization enables us to provide specific tools for browsing the vocabulary, querying the vocabulary or make use of the vocabulary in other ways. In the initial development phase our focus lay on the more common RDF-related stacks, encompassing ontologies in OWL and vocabularies in SKOS. To cater to the diverse needs within these stacks, we have curated a suite of specific RDF tooling:

- **SPARQL:** All RDF distributions are not just cached, but also stored in a SPARQL store and thereby allowing users to query the vocabularies. Each version is stored within its own graph.
- **pyLode:** A documentation tool[23] that understands OWL ontologies. The generated documentation is made available through the static file server.
- **Skosmos:** With Skosmos[24] we have an intuitive browser designed specifically for navigating SKOS vocabularies. The workers provide the necessary configuration for Skosmos to extract the SKOS data from the SPARQL store.
- **Summarizer:** The summarizer leverages the RDFlib library[25] to provide statistics on the RDF distribution. These statistics encompass details such as the namespaces and prefixes defined, the specific entities and classes defined and the languages utilized. This supplementary information offers insights into the usage of other vocabularies and enhances understanding regarding the availability of the vocabulary.

4.3 Recommendations

Recommendations are envisioned in two ways: (1) recommend entire vocabularies based on quality criteria and FAIR assessments; (2) facilitate discovery of individual properties (e.g, from certain schemas) using column names entered by the users. In this way, users can align their own schemas to existing

FAIR schemas. In the initial development phase the registry mainly functions as a metadata catalogue for vocabularies with some addons to cache and visualise them. Although this does help users to find relevant vocabularies, it does not yet help to choose among alternatives. To be able to rank vocabularies, an impression of their community recommendation status should be acquired. To do so the following inputs are taken into account:

- appearance in other catalogues, e.g., LOV[26], or lists, e.g. Awesome Humanities;
- in which fraction of datasets that are harvested for the community is the vocabulary used;
- reviews of the vocabulary by the community

As the registry is still young, some of these inputs are only patchy available, e.g., reviews. But we can already experiment with letting these recommendations steer the output of the vocabulary recommender mentioned before.

5 Self Assessment

To determine the maturity of our vocabulary registry we have taken the model for catalogues of semantic artefacts, which is the outcome of the extensive analysis done by a group within the EOSC Task Force on semantic interoperability [Corcho et al., 2023], to do a self assessment.

- **Metadata (Me)** The vocabulary records are for the most part described using standardized vocabularies, primarily relying on DCAT[27]. While DCAT also describes various distributions/versions of vocabularies, additional metadata, such as reviews and summarized RDF results, utilize Schema.org[28] and VoID[29] respectively. A user interface harmonizes the presentation, alongside accessibility via API and SPARQL.
- **Openness (Op)** The software is fully open source and available for public. The model is also openly available. Registered users can propose new semantic artefacts for inclusion.
- **Quality (Qu)** A pool of maintainers/curators oversee the upkeep of the semantic artifacts, including acceptance or rejection of new submissions. The system generates additional metadata where feasible.
- **Availability (Av)** The registry is freely available without restrictions, though the submission of new semantic artefacts and review functionality is reserved for registered users only. At the moment the primary language of the metadata is English, so multilinguality is an area of improvement.
- **Statistics (St)** The summarizer provides metrics on each semantic artefact and even though we provide some basic catalogue statistics, there is certainly room for improvement there. The review functionality provides the social metrics for the semantic artefacts.
- **PID (Pi)** Persistent identifiers are not currently utilized, necessitating further investigation.
- **Governance (Go)** Governance is overseen by CLARIAH, although the rules of these semantic artefacts are yet to be formalized.
- **Community (Co)** The CLARIAH and CLARIN communities can not only obtain information from the registry, but can write reviews and submit new semantic artefacts for inclusion.
- **Sustainability (Su)** The registry is available through CLARIAH as an organisation, community and a project.
- **Technology (Te)** The registry is available through a web search GUI, but also accessible through an API and through SPARQL. Alignment of semantic artefacts is also an area to investigate further.
- **Transparency (Tr)** The documentation of the data flow and the curation process is still work in progress. Records of previous version are recorded and available.
- **Assessment (As)** This self-assessment is the first assessment of the registry.

6 Future Work and Conclusions

The current implementation still has a RDF focus, i.e. it can deal with SKOS and OWL. However, in the followup of the Dutch CLARIAH project, i.e. SSHOC-NL, which started in the beginning of 2024, it is our task to extend the support beyond RDF into the XML and relational databases world. We also have been working on prototypes that include adding ”memento”[30] support to the SKOSMOS API to ’time travel’ through versions of a SKOS vocabulary and a proxy that allows to locally resurrect domains to make dead links resolvable again. These prototypes can become full implementations to keep the FAIRness of older semantic artefacts. Although the CLARIAH FAIR Vocabulary Registry is still young, it forms the basis of making semantic artefacts relevant for the CLARIN and CLARIAH communities available in a FAIR manner. It already allows resource creators to find existing vocabularies and reuse those; but it is still in a testing phase, which we expect to to move into a stable version that can incrementally be improved. In what concerns the content selection, we initially populated the registry with vocabularies selected by two initiatives: ”Yet Another LOD Cloud” (YALC)[31] and ”CLARIAH Awesome Ontologies for Digital Humanities”[32], but once the testing phase has passed, we will do a careful selection based on the needs of the user group(s) and on the vocabularies already used in previous projects by the CLARIAH and CLARIN communities. The FAIR principles often have a high Linked Data focus, while many research communities have a wealth of information available using other paradigms and related technology. The aim of the CLARIAH FAIR Vocabulary Registry is to also support the semantic artefacts for these resources and make them FAIR.

References

Corcho, O., Ekaputra, F. J., Heibi, I., Jonquet, C., Micsik, A., Peroni, S., & Storti, E. (2023). A maturity model for catalogues of semantic artefacts. *arXiv:2305.06746*. https://arxiv.org/abs/2305.06746

Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2008). Isocat: Corralling data categories in the wild. *6th International Conference on Language Resources and Evaluation*.

McBride, B. (2004). The resource description framework (rdf) and its vocabulary description language rdfs. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 51–65). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24750-0_3

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Windhouwer, M., & Schuurman, I. (2014). Linguistic resources and cats: How to use isocat, relcat and schemacat. *Ninth International Conference on Language Resources and Evaluation*.

Notes

1. clariah.nl
2. dev instance: registry.vocabs.dev.clariah.nl
3. www.w3.org/RDF/
4. go-fair.org/how-to-go-fair/fair-implementation-profile/
5. www.w3.org/OWL/
6. www.w3.org/2004/02/skos/
7. lov.linkeddata.es
8. archivo.dbpedia.org
9. rdamsc.bath.ac.uk
10. www.dcc.ac.uk/guidance/standards/metadata/list
11. metadataregistry.org
12. fairsharing.org
13. ontoportal.org
14. bartoc.org
15. concepts.clarin.eu/ccr
16. www.w3.org/XML/Schema
17. en.wikipedia.org/wiki/Data_definition_language
18. www.w3.org/TR/tabular-data-primer/
19. www.clarin.eu/cmdi
20. docs.celeryq.dev
21. www.w3.org/TR/sparql11-overview/
22. nginx.org
23. github.com/RDFLib/pyLODE
24. skosmos.org
25. rdflib.readthedocs.io
26. lov.linkeddata.es
27. www.w3.org/TR/vocab-dcat-3/
28. schema.org
29. www.w3.org/TR/void/
30. datatracker.ietf.org/doc/html/rfc7089
31. triplydb.com/Triply/yalc
32. github.com/CLARIAH/awesome-humanities-ontologies

Managing Access to Language Resources in a Corpus Analysis Platform

Eliza Margaretha Illig
Department of Digital Linguistics
IDS Mannheim, Germany
margaretha@ids-mannheim.de

Nils Diewald
Department of Digital Linguistics
IDS Mannheim, Germany
diewald@ids-mannheim.de

Paweł Kamocki
Department of Digital Linguistics
IDS Mannheim, Germany
kamocki@ids-mannheim.de

Marc Kupietz
Department of Digital Linguistics
IDS Mannheim, Germany
kupietz@ids-mannheim.de

Abstract

Corpus query tools are crucial to CLARIN's mission of facilitating the sharing and using language data for research. It is a huge challenge for online corpus platforms to manage user access rights for large corpora with complex licenses and heterogeneous restrictions on access methods and purposes. This paper presents an approach to maximize user access to corpus data while protecting rights holders' legitimate interests. Query rewriting techniques and authorization procedures allow for modelling license terms in details, enabling broader applications. This offers an alternative to methods that only model a greatest common denominator of licenses, thereby limiting the possibilities for using the data. Our approach constitutes a flexible and extensible corpus license and user rights management component applicable for other language research environments.

1 Introduction

CLARIN and linguistics in general faces the challenge that its research data are typically affected by the rights of third parties. One approach is to employ technical measures that protects rights holders while minimizing data use restrictions. This is typically done using an online corpus query system allowing only indirect data access. Provided that uniform licenses are available, corpus concordancers, for example, allow authenticated users who have agreed to their terms of service to view keywords in context (KWICs) without allowing full-text reconstruction. The situation becomes more difficult when – as is often unavoidable – different licenses and rights exist for different parts of the data and different groups of users, or when close reading of KWICs is not the only use case. Our paper presents the KorAP (Diewald et al., 2016) approach to making very large corpora, such as the German Reference Corpus DeReKo (Kupietz & Lüngen, 2014), which is affected by more than 200 partly heterogeneous licenses and is used in very different contexts, as usable as possible, while safeguarding the legitimate interests of rights holders.

1.1 Corpus Licenses

Licensing agreements define by whom, for what purposes and how a resource can be used. They can be divided into those limited to "academic" uses, and those allowing commercial uses. For example, they may allow access only to users affiliated with a research organization, with an authenticated account (this also applies to automated access on behalf of the user). Further restrictions may include access only via a dedicated platform, API, from a specific physical location or via a specific network.

Popular public licenses, such as Creative Commons (CC), do not discriminate between groups of users allowing resources to be available on online corpus analysis platforms without authentication. However, even CC licenses contain restrictions on re-use, ranging from simple acknowledgement of the source (BY), to the prohibition of any commercial use (NC).

Statutory exceptions in applicable copyright legislation can also be interpreted as 'licenses' *sui generis*, i.e. permissions granted not by the rights holder, but directly by the legislator. The statutory exceptions for Text and Data Mining (TDM), harmonized at the EU level by the Digital Single Market (DSM) Directive (2019/790), deserve special attention. Article 3 of the DSM Directive allows research organisations to build corpora for TDM purposes, which (at least in certain EU jurisdictions such as Germany) can subsequently be shared with research partners.

Certain metadata and annotations can be licensed (even though it seems that by themselves metadata would rarely attract copyright protection, they may still be protected by the *sui generis* database right). Generally, metadata are in the public domain, and can be released and shared without restrictions.

1.2 User Rights Management

Corpus licenses determine which users have which access rights to which parts of primary data, metadata, and annotation data (the latter being determined by software licenses as well). The rights of a user therefore have to be managed in addition by a corpus platform and can be matched with the licenses after authentication and before any data can be delivered to an account. As previously introduced, these licenses determine not only whether, but also in which ways a user can access the data. This requires that access rights must not only be compared statelessly and statically, but also take into account the temporal and local contexts. For example, if licenses only allow short excerpts from texts (e.g., KWICs), there is a reasonable concern of licensors that the original full-text can be reconstructed from the search or analysis results by cleverly formulating follow-up search queries.[1] In order of prohibiting such use, it may be necessary to monitor an account's search queries over time to detect and/or prevent misuse.

Some corpus licensors also limit the availability of their data to sites within a specific location or network. The user rights management of an online corpus platform must be able to match the IP address of the account with the address space allowed for access. If licensors make licenses available to users only for a limited period of time, the user rights management system must log the initial access and check with each access whether the approved time frame has not yet been exceeded.

2 Related Work

In digital rights management, Open Digital Rights Language (ODRL; Iannella and Villata, 2018) is commonly used to represent policies on the usage of digital content and services. While ODRL focuses constraints on parties, assets or actions, our approach emphasizes constraints based on licenses. For authentication and authorization, Shibboleth (Cantor & Scavo, 2005) facilitates Single Sign-On (SSO) typically used by academic users to access corpora with academic licenses provided by corpus platforms e.g. OpenSoNaR (Reynaert et al., 2014). Shibboleth, as well as OpenID and LDAP, only offer limited and static user configurations for authentication, thus does not cover all access control requirements in KorAP.

Keycloak (Thorgersen & Silva, 2021) offers more control and management of authentication and authorization by providing a comprehensive admin console, e.g. to customize an authentication flow. The BlackLab (de Does et al., 2017) corpus search engine is integrating Keycloak. Google Zanzibar (Pang et al., 2019) features a distributed authorization system for managing user access across numerous applications rapidly. While Keycloak and Google Zanzibar manage access based on user roles and groups, KorAP also requires access control based on licenses.

3 KorAP Approach

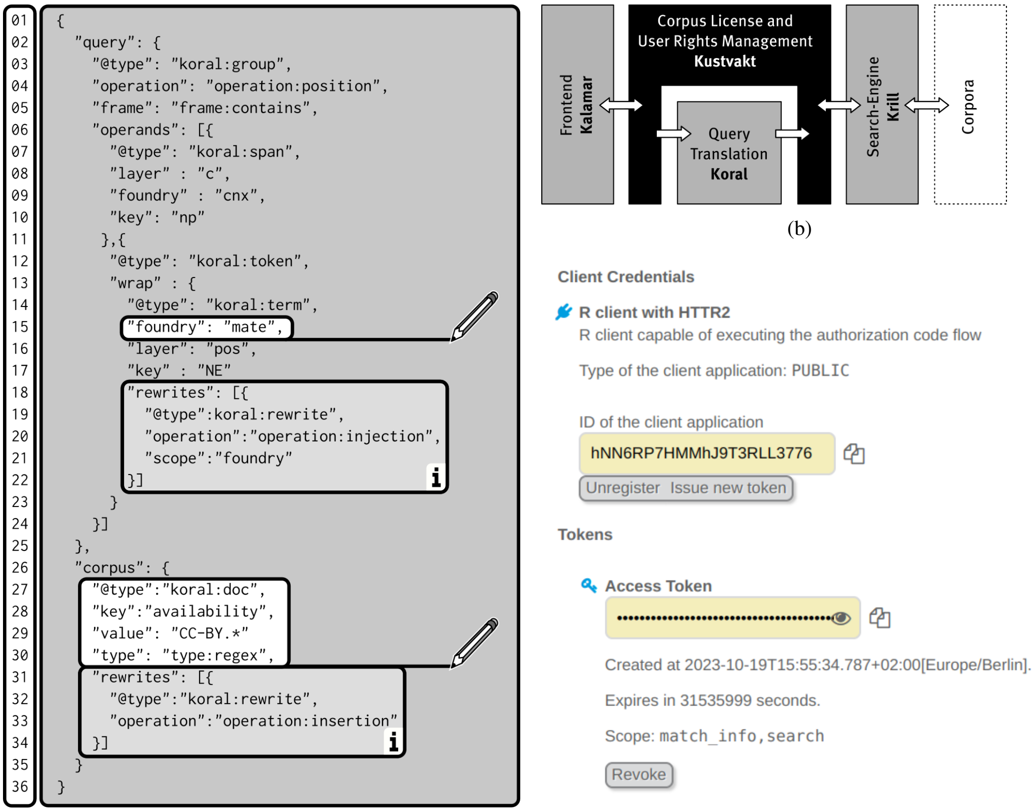
The challenge of a corpus license and user rights management system is to find technical solutions for mapping rights and licenses and restricting data access accordingly. To maintain flexibility and independence from underlying data and user interfaces, the rights management in KorAP is a separate server-based component called Kustvakt (Margaretha & Contributors, 2023)[2]. It receives (authorized) API requests, e.g. from the web UI Kalamar or other clients, and uses Koral to translate queries to *KoralQueries* (Bingel & Diewald, 2015). Kustvakt performs query rewriting on KoralQueries respecting user rights, forwards them to the search engine Krill and returns their responses to the requesting entity (see Fig. 1b).

3.1 Query Rewrites

To manage access to a resource in terms of both licenses and user rights while granting the user the greatest possible amount of liberty, an approach based on *query rewriting* was chosen. In this approach, a resource request is reformulated to correspond to the access rights and can be answered by the database

[1] This is also a concern for corpus use in language models.

[2] https://github.com/KorAP/Kustvakt



```
01 {
02   "query": {
03     "@type": "koral:group",
04     "operation": "operation:position",
05     "frame": "frame:contains",
06     "operands": [{
07       "@type": "koral:span",
08       "layer" : "c",
09       "foundry" : "cnx",
10       "key": "np"
11      },{
12       "@type": "koral:token",
13       "wrap" : {
14         "@type": "koral:term",
15         "foundry": "mate",
16         "layer": "pos",
17         "key" : "NE"
18         "rewrites": [{
19           "@type":koral:rewrite",
20           "operation":"operation:injection",
21           "scope":"foundry"
22         }]
23       }
24     }]
25   },
26   "corpus": {
27     "@type":"koral:doc",
28     "key":"availability",
29     "value": "CC-BY.*"
30     "type": "type:regex",
31     "rewrites": [{
32       "@type":"koral:rewrite",
33       "operation":"operation:insertion"
34     }]
35   }
36 }
```

(a)

(b)

Client Credentials

R client with HTTR2

R client capable of executing the authorization code flow

Type of the client application: `PUBLIC`

ID of the client application

hNN6RP7HMMhJ9T3RLL3776

Tokens

Access Token

••••••••••••••••••••••••••••••••••••

Created at 2023-10-19T15:55:34.787+02:00[Europe/Berlin].

Expires in 31535999 seconds.

Scope: `match_info,search`

(c)

Figure 1: (a) The corpus query "Return all nominal phrases as annotated in the `cnx` foundry that contain a named entity" is rewritten to use a default annotation for part-of-speech and restricted to all corpora licensed under Creative Commons. (b) The corpus and user management component is an independent middleware broker service between API requests and the search engine (c) OAuth client and token management in the web UI

without further knowledge of licenses and user rights (cf. Rizvi et al., 2004). To achieve this, restrictions in the form of metadata are encoded directly at the individual text level and can thus be excluded directly during a search or analysis on the corpora. In principle, it is possible to take any metadata into account in the rewrite process, for example the identification of a license (as in Fig. 1a via the metadata field `availability`, line 28), but also corpus labels or author names. The strategy is fundamentally *additive*, which means that restrictions are lifted by adding rules, i.e. the user receives a more extensive query with more authorizations, which allows for greater access. The basis for this query rewriting is KoralQuery an implementation of CQLF (Bański et al., 2016) in form of a unified JSON-LD-based (Sporny et al., 2014) representation of an abstract corpus query that is independent of any corpus query language (such as CQP, Annis-QL, Poliqarp, and Cosmas-2-QL). By adding or changing constraints (e.g. restricting a virtual corpus to all texts with a CC license), a new query can be formulated that satisfies all requirements and can be passed to the database.

In addition to restrictions on text access, it is also possible to exclude query options. Rules can be formulated that exclude the search in certain annotations or set defaults for queries on annotations (e.g., see Fig. 1a, line 15). Any modification to the query is marked (see Fig. 1a, lines 18–22 and 31–34). This may be necessary as it is the only way to provide feedback to the user as to which resources they requested they actually have access to.

Query rewriting is independent of the corpus and the user size, therefore it scales and performs well with a growing database. It is only dependent on the different restrictions that need to be lifted in the case of authorizations granted to the user.

3.2 Access policies

To access most of DeReKo, users must agree to our terms of use during user-registration to use KorAP. We specify 3 types of access policies based on login and access location: 1. Free access on corpora under CC licenses accessible from anywhere without login, e.g. Wikipedia; 2. Public access on free and academic corpora that requires login; 3. All corpora access requiring login and access through our network. By login, we mean not only user authentication but also authorization given to a third-party application (see Section 3.3). We use LDAP for authentication and IP ranges to determine access location. DeReKo is annotated with an `availability` metadata field representing their licenses with categorizations, e.g. CC, ACA (academic), and QAO-NC (query-analysis-only, non-commercial) introduced in Kupietz and Lüngen, 2014. We use it to apply the access policies by using query rewriting (see Section 3.1).

To prevent a reconstruction of original texts from search results as described in Section 1.2, we impose a limit to the size of the match context. Besides, we employ a timeout mechanism to restrict search duration thereby enhancing system responsiveness.

3.3 Authorization

KorAP supports the authorization framework OAuth 2.0 (Hardt, 2012) defining communication protocols with client applications and granting authorizations called access tokens. Kustvakt provides web-service APIs and acts as an authorization server issuing and managing access tokens (Kupietz et al., 2020). We support the authorization code grant flow for server-based client applications to obtain access tokens. For non-server-based clients such as desktop applications incapable of handling HTTP requests, we provide a feature to obtain access tokens from our web UI Kalamar as shown in Figure 1c. It is also possible to perform the authorization code flow by using local web-servers or libraries, e.g. RKorAPClient (Kupietz et al., 2020) uses httr2 (Wickham, 2023).

OAuth 2.0 differentiate clients into 2 types: confidential clients that can keep a client secret thus can authenticate securely, and public clients that cannot. To protect users from authorization abuse by malicious software, we limit the time validity of access tokens depending on these client types. Confidential clients receive short-live access tokens and a long-live refresh token to request a new access token without re-authorization. Public clients receive long-live access tokens without refresh token. Additionally, KorAP requires client registration and provides token revocation (Lodderstedt & Scurtescu, 2013) via API or web UI. Kalamar acts as a frontend to the API and facilitates client and token management by listing all registered clients and their access tokens. Furthermore, users can revoke access tokens when suspecting misuse or delete clients as described in Figure 1c.

4 Extensibility

KorAP has already covered all access rights requirements for DeReKo. Some extensions can be profitable as follows. While metadata is generally freely available regardless of access restriction on corpus content, some fields can be restricted based on licensing limiting access to them. In addition to policy enforcement, the protocol-based approach enables the integration of other query rewriting methods, such as query expansion (cf. Baeza-Yates & Ribeiro-Neto, 2010, ch. 5) independent of the user and corpus base. This approach allows applying cascading rewrites to a query with policy enforcement at the end to prevent unintended permission expansion. Moreover, *response rewriting* can be performed to filter a result set according to certain criteria before returning it to an account. Since response rewriting usually requires requesting excess data from the resource, it is only suitable for small result sets, e.g. for shortening of text snippets (when certain text licenses allow longer contexts than others).

5 Conclusion

Directly integrating policy enforcement at the protocol level through query rewriting and abstract authorization mechanisms allows for a great deal of transparency and flexibility for efficient and detailed access control to corpus resources with complex licenses. Our approach facilitates maximum access and usage of corpora while ensuring compliance with complex licenses. The currently applied rule set in our implementation is based on the needs of the different licenses of DeReKo, so the full flexibility is not yet exhausted. The largest application using our approach is currently a corpus query system, serving a corpus of 87 million texts for an average of 6000 queries per day. Kustvakt is open source and in conjunction with KoralQuery universally applicable for resource control in corpus analysis applications.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (2010). *Modern Information Retrieval: The Concepts and Technologies behind Search* (2nd ed.). Addison-Wesley.

Bański, P., Frick, E., & Witt, A. (2016). Corpus Query Lingua Franca (CQLF). *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 2804–2809.

Bingel, J., & Diewald, N. (2015). KoralQuery - a General Corpus Query Protocol. *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015.*

Cantor, S., & Scavo, T. (2005). Shibboleth architecture. *Protocols and Profiles*, *10*(16), 29.

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with blacklab. *CLARIN in the Low Countries*, 245–257.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Bański, P., & Witt, A. (2016). KorAP architecture - Diving in the Deep Sea of Corpus Data. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 3586–3591.

Hardt, D. (2012, October). The OAuth 2.0 Authorization Framework. https://doi.org/10.17487/RFC6749

Iannella, R., & Villata, S. (2018). ODRL Information Model 2.2. *W3C Recommendation*, *15*.

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., & Lindström, N. (2014). JSON-LD 1.0. A JSON-based Serialization for Linked Data. http://www.w3.org/TR/json-ld/

Kupietz, M., Diewald, N., & Margaretha, E. (2020). RKorAPClient: An R package for accessing the German Reference Corpus DeReKo via KorAP. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC '20)*, *12*, 7015–7021.

Kupietz, M., & Lüngen, H. (2014). Recent developments in DeReKo. *Proceedings of the 9th conference on international language resources and evaluation (LREC'14)*, 2385.

Lodderstedt, T., & Scurtescu, M. (2013, August). *OAuth 2.0 Token Revocation* (RFC No. 7009). RFC Editor. https://tools.ietf.org/html/rfc7009

Margaretha, E., & Contributors. (2023, September). *KorAP/Kustvakt: version 0.71.1*. Zenodo. https://doi.org/10.5281/zenodo.8389644

Pang, R., Caceres, R., Burrows, M., Chen, Z., Dave, P., Germer, N., Golynski, A., Graney, K., Kang, N., Kissner, L., Korn, J. L., Parmar, A., Richards, C. D., & Wang, M. (2019). Zanzibar: Google's Consistent, Global Authorization System. *2019 USENIX Annual Technical Conference*.

Reynaert, M., van de Camp, M., & van Zaanen, M. (2014). OpenSoNaR: User-driven development of the SoNaR corpus interfaces. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, 124–128.

Rizvi, S., Mendelzon, A., Sudarshan, S., & Roy, P. (2004). Extending query rewriting techniques for fine-grained access control. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 551–562. https://doi.org/10.1145/1007568.1007631

Thorgersen, S., & Silva, P. I. (2021). *Keycloak-identity and access management for modern applications: harness the power of Keycloak, OpenID Connect, and OAuth 2.0 protocols to secure applications*. Packt Publishing Ltd.

Wickham, H. (2023). *httr2: Perform HTTP Requests and Process the Responses* [https://httr2.r-lib.org, https://github.com/r-lib/httr2].

IcePaHC 2024.03 – A Significant Treebank Upgrade

Joel C. Wallenberg
University of York
joel.wallenberg@york.ac.uk

Anton Karl Ingason
University of Iceland
antoni@hi.is

Einar Freyr Sigurðsson
Árni Magnússon Institute for Icelandic Studies
einar.freyr.sigurdsson@arnastofnun.is

Eiríkur Rögnvaldsson
University of Iceland
eirikur@hi.is

Abstract

The version of the Icelandic Parsed Historical Corpus (IcePaHC) that was released in 2011 and later made available through CLARIN has facilitated a wide range of studies, both in terms of theoretical linguistics and Natural Language Processing. Here, we discuss how IcePaHC has been used throughout the years and present the first major update to IcePaHC in 13 years, version 2024.03, involving thousands of corrections that allow for more precise research than before. The current version is released under a Creative Commons Attribution license (CC BY).

1 Introduction

The Icelandic Parsed Historical Corpus (IcePaHC) (Rögnvaldsson et al., 2011, 2012; Wallenberg et al., 2011) is a manually annotated phrase structure treebank that contains approximately 1 million words of parsed text that has been sampled from historical data.[1] The treebank contains texts from every century from the 12th century to the 21st century, mostly from narratives and religious texts that are spread more or less evenly across this period.

When IcePaHC 0.9 was released in 2011, it was the first major treebank for Icelandic and thus it paved the way for various types of studies that had not been feasible before. However, a treebank is a complicated dataset that must be maintained because there are no practical methods available that make sure that a treebank is completely free from errors. Despite various manual and automatic methods to minimize errors, it remains an ongoing process to fix mistakes in the data, and thus the GitHub repository for IcePaHC (https://github.com/antonkarl/icecorpus/) has evolved continuously over the last 13 years, and especially during 2023 and 2024 as the authors of the corpus have made a systematic effort to correct as many errors as possible. The outcome of these efforts is the release of a new major update of the treebank to CLARIN, version 2024.03, now available for download (Wallenberg et al., 2024).

This paper is organized as follows. Section 2 reviews some background on the IcePaHC treebank and related resources. In Section 3 we discuss how the corpus has been used in linguistics research and in Section 4 we go over some studies that have been carried out using IcePaHC that involve Natural Language Processing (NLP). Section 5 discusses the new version and Section 6 concludes.

2 Background

The IcePaHC corpus has its roots in a research program that goes back to the annotation of the Penn Treebank, the first major phrase structure treebank (Marcus et al., 1993). The Penn Treebank was, in turn, followed by the Penn Parsed Corpora of Historical English, also developed at the University of Pennsylvania (Kroch & Taylor, 2000b; Kroch et al., 2004). This second iteration of treebank development involved some improvements to the annotation scheme, including a more flat structure in cases that involved ambiguity, such as in PP-attachment and the ordering of elements within the verb phrase. This is important in order to not have the annotation involve too many decisions that are simply based on a

[1] We thank three anonymous reviewers for their comments on this paper.

convention rather than a reliable analysis and also because during historical change, the theoretically appropriate analysis is not always clear as grammar competition may be ongoing in the speech community (Kroch, 1989, 2001; Kroch & Taylor, 2000a).

This tradition of annotating historical corpora was continued and extended to the Icelandic language in the project Viable Language Technology beyond English – Icelandic as a test case whose PI was Eiríkur Rögnvaldsson. This served a dual function as the treebank was intended both for linguistic research and development of Natural Language Processing tools (Rögnvaldsson, 2010). The IcePaHC project used a semi-automatic method, for example by running the shallow parser IceParser (Loftsson & Rögnvaldsson, 2007) and structure-modifying search queries in CorpusSearch (Randall, 2005) before manual correction and further manual annotation of the phrase structure was carried out.

3 Uses in Linguistics

The IcePaHC treebank has been used for several linguistics studies. For example, Ingason et al. (2013) used the frequencies of certain types of passive constructions to model the evolutionary trajectory of the Icelandic New Passive and make predictions about its spread, even into the future. In this case, it was the Variational Model of Language Acquisition (Yang, 2002) that provided a theoretical foundation for a predictive analysis but the treebank was the source of the empirical counts that were used in the analysis.

Wallenberg et al. (2021) used IcePaHC to study the relationship between information smoothness in the sense of Information Theory and the trajectory of diachronic change. In this case, word orders that serve as a diagnostics for certain constructions were extracted along with features that characterized the environment of each example and the text of each sentence was submitted to a function that calculated an indicator of information smoothness/density. The study found that Information Theoretic factors have a significant effect on how historical change evolves over time.

Various other topics have been studied, including expletives and cataphora (Booth, 2018, 2019), and verb second vs. verb first word orders (Booth & Beck, 2021). Schätzle (2018) furthermore studied dative subjects in the history of Icelandic with an emphasis on data visualization.

4 Uses in Natural Language Processing

The most prominent task for treebanks in Natural Language Processing is the possibility of training data-driven parsers on the manually annotated trees such that a system becomes available for automatic parsing of the same type of trees, given any arbitrary text. IcePaHC was used along with the Faroese Parsed Historical Corpus (FarPaHC) (Ingason et al., 2012) for an experiment that made use of the fact that the two insular Scandinavian languages have a similar syntax. The experiment involved training a parser on a mixture of the two languages and found that parsing accuracy in Faroese can be improved by adding Icelandic data to the training dataset (Ingason et al., 2014).

A pipeline for parsing Icelandic text was trained and made available by Jökulsdóttir et al. (2019). This CLARIN resource focused on providing the basic infrastructure needed for setting up parsing pipelines for Icelandic and it included a configuration of the Berkeley parser that had been trained on IcePaHC. This pipeline setup was used to facilitate the first release of a neural parsing pipeline for Icelandic in Arnardóttir and Ingason (2020), a system that got an F1 score of 84.74% and was also trained on IcePaHC – and furthermore it was supported by the application of a multilingual BERT model. Although IcePaHC is a phrase structure treebank, it has also served as the foundation for development of dependency parsing for Icelandic due to the development of conversions from phrase structure to Universal Dependencies (UD) (Arnardóttir et al., 2020; Arnardóttir et al., 2023). Both IcePaHC and FarPaHC have been converted to a UD format. Furthermore, IcePaHC served as a model for the parsing of parliament speeches (Rúnarsson & Sigurðsson, 2020) and sports-news texts[2] with the parsing subsequently being converted to a UD format. The newest versions of all three UD treebanks, UD_Icelandic-IcePaHC (Arnardóttir, Hafsteinsson, Sigurðsson, Jónsdóttir, et al., 2024), UD_Icelandic-Modern (Rúnarsson et al., 2024) and

[2] The parsed texts have not been published as of yet but the parsing, carried out by Kristján Rúnarsson, can be found at our GitHub repository: https://github.com/antonkarl/icecorpus/.

UD_Faroese-FarPaHC (Arnardóttir, Hafsteinsson, Sigurðsson, Ingason, et al., 2024), are found in the latest release of Universal Dependencies.

5 The New Version

Building on the success of IcePaHC and its open access release on CLARIN, we now present a new major upgrade of the resource. The treebank contains the same texts but thousands of corrections have been made to the annotation, resulting in a more accurate resource for use in both linguistics and language technology. The new and updated version has already been made available on CLARIN and it is called IcePaHC 2024.03. This means that instead of using version numbers like 0.3, 0.5, 0.9, like we have done in the past, the version number now follows a system where the release date is used as the basis of the numbering, March 2024 in this case.

Various types of corrections have been made to both syntactic structure, Part-of-Speech tags and lemmatization. This last point is particularly significant since thousands of corrections involve correcting lemmas in the corpus. This is particularly useful when designing queries that target the dictionary form of a word. Structure correction has also made use of the fact that the treebank is in open access and therefore enjoys regular feedback from its users. We are grateful for the emails we have received about aspects of the annotation that needed to be reconsidered and we have done so on many occasions throughout the years.

6 Conclusion

We have described the IcePaHC corpus, its many uses, and the new significant upgrade that has now been released. While several language resources have been made available in the last few years (Nikulásdóttir et al., 2022), Icelandic remains a low-resource language (Rehm & Way, 2023) and therefore every step counts along the path towards a more robust Language Technology ecosystem for the language. The release of IcePaHC 0.9 was a major step in 2011 and facilitated diverse research in the following years. Now that a new version has been made available with more precise annotation, we remain optimistic that the tradition of studying Icelandic phrase structure computationally continues to be a fruitful enterprise.

References

Arnardóttir, Þ., Hafsteinsson, H., Jasonarson, A., Ingason, A., & Steingrímsson, S. (2023). Evaluating a Universal Dependencies Conversion Pipeline for Icelandic. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 698–704.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., & Steingrímsson, S. (2020). A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, 16–25.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Ingason, A. K., Rögnvaldsson, E., & Wallenberg, J. C. (2024). UD_Faroese-FarPaHC. In Universal Dependencies 2.14.

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Jónsdóttir, H., Bjarnadóttir, K., Ingason, A. K., Rúnarsson, K., Steingrímsson, S., Wallenberg, J. C., & Rögnvaldsson, E. (2024). UD_Icelandic-IcePaHC. In Universal Dependencies 2.14. http://hdl.handle.net/11234/1-5502

Arnardóttir, Þ., & Ingason, A. K. (2020). A Neural Parsing Pipeline for Icelandic Using the Berkeley Neural Parser. *Proceedings of CLARIN Annual Conference*, 48–51.

Booth, H. (2018). *Expletives and Clause Structure. Syntactic Change in Icelandic* [Doctoral dissertation, The University of Manchester].

Booth, H. (2019). Cataphora, expletives and impersonal constructions in the history of Icelandic. *Nordic Journal of Linguistics*, *42*(2), 139–164.

Booth, H., & Beck, C. (2021). Verb-second and verb-first in the history of Icelandic. *Journal of Historical Syntax*, *5*(28), 1–53.

Ingason, A. K., Legate, J. A., & Yang, C. (2013). The Evolutionary Trajectory of the Icelandic New Passive. *University of Pennsylvania Working Papers in Linguistics*, *19*(2), 91–100.

Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Wallenberg, J. C. (2014). Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 91–95.

Ingason, A. K., Sigurðsson, E. F., Rögnvaldsson, E., & Wallenberg, J. C. (2012). Faroese Parsed Historical Corpus (FarPaHC) 0.1 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/92

Jökulsdóttir, T. F., Ingason, A. K., & Sigurðsson, E. F. (2019). A Parsing Pipeline for Icelandic Based on the IcePaHC Corpus. *Proceedings of CLARIN Annual Conference*, 138–141.

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*, 199–244.

Kroch, A. S. (2001). Syntactic Change. In M. Baltin & C. Collins (Eds.), *The Handbook of Contemporary Syntactic Theory* (pp. 698–729).

Kroch, A. S., Santorini, B., & Delfs, L. (2004). Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition. [Size: 1.8 million words.]

Kroch, A. S., & Taylor, A. (2000a). Verb-Object Order in Early Middle English. *Diachronic Syntax: Models and Mechanisms*, 132–163.

Kroch, A. S., & Taylor, A. (2000b). Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition. [Size: 1.3 million words.]

Loftsson, H., & Rögnvaldsson, E. (2007). IceParser: An Incremental Finite-State Parser for Icelandic. *Proceedings of the 16th Nordic Conference of Computational Linguistics*, 128–135.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, *19*(2), 313–330.

Nikulásdóttir, A. B., Arnardóttir, Þ., Barkarson, S., Guðnason, J., Gunnarsson, Þ. D., Ingason, A. K., Jónsson, H. P., Loftsson, H., Óladóttir, H., Rögnvaldsson, E., Sigurðsson, E. F., Sigurgeirsson, A. Þ., Snæbjarnarson, V., Steingrímsson, S., & Örnólfsson, G. T. (2022). Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS. *Selected Papers from the CLARIN Annual Conference 2021*, 109–125.

Randall, B. (2005). CorpusSearch 2 User's Guide. University of Pennsylvania.

Rehm, G., & Way, A. (Eds.). (2023). *European Language Equality – A Strategic Agenda for Digital Language Equality*.

Rúnarsson, K., & Sigurðsson, E. F. (2020). Parsing Icelandic Alþingi Transcripts: Parliamentary Speeches as a Genre. *Proceedings of the Second ParlaCLARIN Workshop*, 44–50.

Rúnarsson, K., Arnardóttir, Þ., Hafsteinsson, H., Barkarson, S., Jónsdóttir, H., Steingrímsson, S., & Sigurðsson, E. F. (2024). UD_Icelandic-Modern. In Universal Dependencies 2.14.

Rögnvaldsson, E. (2010). Icelandic language technology: An overview. *Language, Languages and New Technologies: ICT in the Service of Languages*, 187–195.

Rögnvaldsson, E., Ingason, A. K., & Sigurðsson, E. F. (2011). Coping with Variation in the Icelandic Parsed Historical Corpus (IcePaHC). *Language Variation Infrastructure*, 97–112.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., & Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1977–1984.

Schätzle, C. (2018). *Dative subjects: Historical change visualized* [Doctoral dissertation, University of Konstanz].

Wallenberg, J. C., Bailes, R., Cuskley, C., & Ingason, A. K. (2021). Smooth Signals and Syntactic Change. *Languages*, *6*(2), 60.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). Icelandic Parsed Historical Corpus (IcePaHC) 0.9 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/62

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2024). Icelandic Parsed Historical Corpus (IcePaHC) 2024.03 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/325

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford University Press.

Towards a Swedish Sign Language Dataset With Pose Estimation Information: Process and Challenges

Gustaf Gren
Department of Linguistics
Stockholm University
gustaf.gren@ling.su.se

Abstract

The Swedish Sign Language Corpus (SSLC) is an ongoing research sign language project, currently sitting at 598 video files of which 261 are available to the public, annotated with sign glosses and translations of sentences into Swedish. The publicly available video files contain about 8600 unique signs and 75,000 tokens. In this paper, we discuss the process and challenges of adding pose estimation information to the corpus, given a partially unstructured collection of raw video data. The resulting dataset contributes a rare type of publicly available resource to the field of computational sign language research, including continuous features derived directly from video data.

1 Introduction

Computational research for sign languages is difficult since most available datasets are either very small, highly domain specific (e.g. from news broadcasts, as in Albanie et al. (2021)), or only include isolated signs (De Sisto et al., 2022). The Swedish Sign Language Corpus (SSLC) (Öqvist et al., 2020) has been developed since 2003 and is a part of the *CLARIN resource family*. The corpus exists in a development version, currently comprising 598 manually annotated video files each containing a conversation between two signers, and a publicly published version comprising 261 files. The latter is available at The Language Archive, part of the CLARIN B-centre Max Plank Institute for Psycholinguistics,[1] as well as at https://teckenspraakskorpus.su.se/.

The aim of this paper is to pave a way to create a dataset using all the data possible from the SSLC which can then directly be applied to sign language research, including sign language recognition and general sign language research. The paper is divided into two distinct sections: i) illustrating the challenges in processing these video files, described in section 2, and ii) discussion about privacy/licensing issues and an introduction and details of pose inferences for this dataset as an additional modality, discussed in section 3 on page 3.

2 Preparing Video Files For Dataset

Each of the 598 video files in the SSLC was extracted out from raw footage of a longer conversation. The raw footage consists of 213 unordered video files in total. Since the video files in the SSLC are of inadequate quality for many research purposes and for making pose estimations, the corresponding video had to be extracted from the raw footage which is of a higher resolution, and later aligned to the SSLC annotations. Since the information on which raw footage contained which SSLC video files was unavailable, an algorithm was devised to identify the raw footage file containing each SSLC video file and compute a frame-by-frame alignment, which in turn could be used to align to the SSLC annotations. This problem was exacerbated by the following inconsistencies in the SSLC data: 1) varying resolutions, 2) intro/outro with differing lengths, 3) accidental cropping (e.g. hands out of frame), 4) varying aspect ratios, and 5) instances of incorrect duplicate frames.

[1] http://hdl.handle.net/11372/DOC-91

While the algorithm was devised for this specific task in hand, the code is dynamic enough to encompass general usecases for matching up unsorted folders of clips and finding out if and where they are situated in original video files.

The general algorithm iterates over each raw video file, and then for each video file in the SSLC still not identified does the following to deal with both the attribution and intro/outro issues:

1. Each SSLC video's offsets (i.e. length of intro/outro) are found. This is done by iterating over each frame, calculating the pixel mean until it finds a frame with a value over a certain boundary (since the mean of a solid black frame is 0). For the intro, iterate from the first frame and forward. For the outro, from the last frame and backwards.
2. A sliding window approach is employed to compare every frame of the SSLC video file with a corresponding window of frames in the original video to make a decision on clip attribution. This involves moving the SSLC video file frame window step-by-step through the original video frames, and is illustrated in figure 1. Each frame in both is normalized by resizing to `160x90` and turned grayscale to make computation cheaper.
    (a) For each position of the window, a distance metric (mean absolute difference) is calculated between the frames of the original and the SSLC video file. This helps in determining how similar the window of frames in the original video is to the frames in the SSLC video file.
    (b) While calculating the distances, if the distance at any point exceeds a predefined early stopping threshold, the algorithm stops further processing for efficiency. This is based on the premise that a very high distance indicates a low probability of the SSLC video file being part of the original video at that point.
3. Once all relevant distances are calculated and none have triggered the early stopping rule, sort the difference array according to distance and calculate the difference between the minimum distance and the *n*th minimum frame. If it meets a threshold then we call that a match.

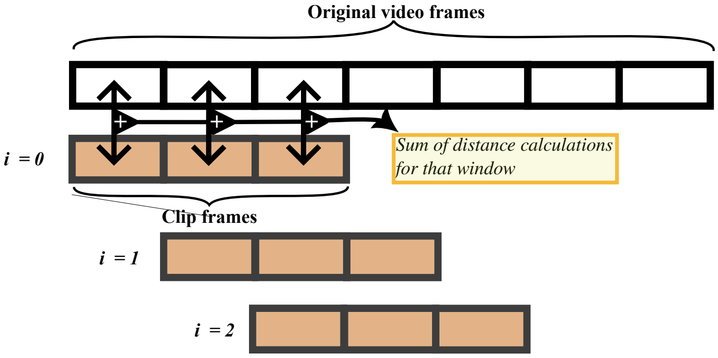
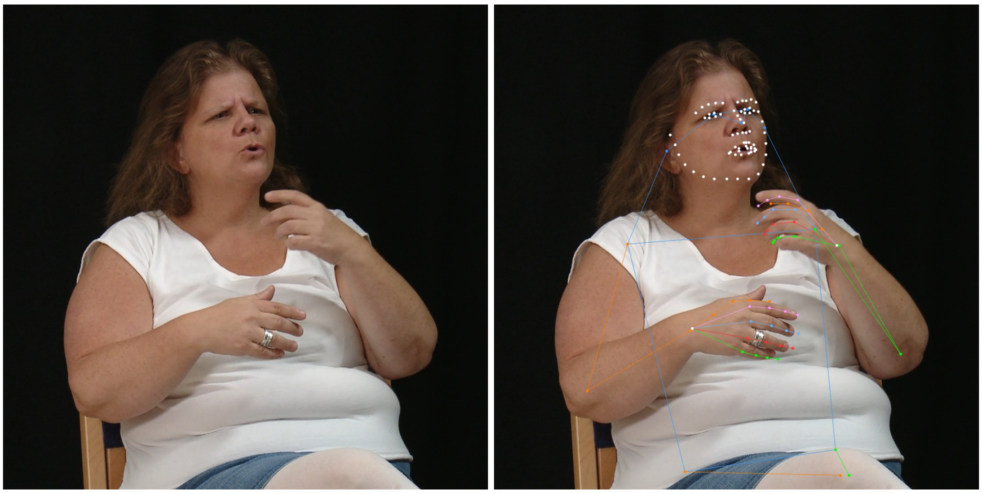


Figure 1: Illustration of sliding window calculation

An initial run with early stopping set intentionally low captured about a quarter of the 598 video files in the SSLC, from which a subset was manually checked to make sure that it correctly extracted the timestamps for each video file as well as the intro/outro length. Since there was no attribution information to go on this was the only means of evaluation available. This process was then repeated with early stopping set progressively higher, which added the count of correctly identified clips and location up to about 86% as of 2024-09-04.

Worth noting is that this algorithm was robust enough to successfully identify clip attribution even in some cases when errors had been made in processing the clips, errors like incorrect cropping and incorrectly stretched aspect ratio.



(a) Raw frame screenshot

(b) Keypoint inferences

Figure 2: Example frame from the dataset, cropped for this paper.

The duplicate frame issue previously mentioned is regarding 33 clips that contain duplicate frames not present in the raw footage. Work is ongoing on an implementation using a dynamic search algorithm to detect the existence of duplicate frames to facilitate their removal, so that the clips can be used as input for the search algorithm described above.

All clips were then extracted and deinterlaced using FFMPEG. Since one cannot trim a video exactly without also transcoding them, they were transcoded into H264 using the *slow* preset and a CRF of 18[2].

For the exact implementation of the algorithm, including the specific parameters for e.g. frame skip and early stopping threshold, I refer to the code on the GitHub repo, available here: `https://github.com/skogsgren/sslc-pose`.

3 Pose Estimation

Keypoint extraction are pairs of x and y coordinates for every frame in a video for a set amount of *keypoints* depending on the specific dataset. The COCA dataset (Jin et al., 2020), for example, has 21 keypoints for each hand. These keypoints allow us to reduce the number of dimensions from a raw video file to a list of vectors with keypoints for how the body, hands, and face move for each frame (for an example, see figure 2).

Datasets with keypoints have been widely utilized in general linguistic research. Östling et al. (2018) utilized keypoints to research iconicity across sign languages. Börstell (2023) observed promise in using pose inferences to estimate the articulation phase of signs. Adding the pose inference modality to the dataset would make this type of research more easily accessible for researchers. Keypoints are also used extensively in the field of machine recognition/translation of sign language (Núñez-Marcos et al., 2023), although there are concerns with using just keypoints without any further adaption for that purpose (Moryossef et al., 2021).

To perform the keypoint inferences the Python library MMPOSE (2020) was used. MMPOSE is a library providing both open source pose detection models and datasets. The `rtmpose-l` model was used, which is a wholebody model based on the models proposed in Jiang et al. (2023) and Lyu et al. (2022). It uses the keypoints as defined in COCO 2020 (Jin et al., 2020). Mediapipe (Lugaresi et al., 2019) is another alternative for keypoint inferences, however it was ultimately not chosen in virtue of MMPOSE's compatibility with the OpenMMLab suite of tools and models.

[2] To maintain as much of the original quality as possible without creating enormous files

4 Availability

The plan is to release the dataset within the CLARIN-network through the same channels as for the current development version of the SSL corpus, that is either through the Language Archive or through the STS-korpus website (see Section 1 on page 1 for links to these). All the data which is matched and appropriately licensed will be released.

5 Discussion and Future Work

Keypoint inferences are not perfect. As we can see in figure 2 on the previous page sometimes the keypoints are not aligning where they should (take the right thumb for example, which should point upwards). When body parts overlap this effect becomes even more prominent (this is discussed in detail in Moryossef et al. (2021)). However, how large of an effect this discrepancy has on performance on various downstream tasks for this particular dataset remains to be seen.

There are many possibilities to expand the dataset further, for example by using the *discrete* data present in the *Swedish sign language lexicon* which contains video clips for $\sim 25,000$ isolated signs, or by anonymizing the remaining parts of the SSLC in order to comply with GDPR. Additionally, since all the clips in the SSLC were recorded using a top-down camera one could also explore adding that dimension although it is unfortunately currently not aligned to the annotations.

6 Conclusion

In this paper we have discussed the challenges in turning the Swedish Sign Language Corpus into a high quality video dataset with pose inferences. This included a novel algorithm for finding clip attribution for unsorted video files, which could be helpful for researchers in the possession of a large quantity of low quality video and unsorted raw footage. While a large portion of the available material remains unavailable to the public due to GDPR, with ~ 8600 unique signs and $75,000$ tokens this dataset has the potential to ease research both within general linguistic inquiries and natural language processing applications for Swedish Sign language.

Acknowledgments

This research was supported by the Swedish national research infrastructure Språkbanken and Swe-Clarin, funded jointly by the Swedish Research Council (2018–2024, grant agreement 2017-00626) and the 10 participating partner institutions. Computations have been facilitated by an Nvidia GPU grant. Special thanks to Robert Östling for being a seemingly endless well of great ideas, and Nikolaus Kankkonen for providing access and help with the data.

References

Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., & Zisserman, A. (2021). Bbc-oxford british sign language dataset.

Börstell, C. (2023). Extracting sign language articulation from videos with MediaPipe. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th nordic conference on computational linguistics (nodalida)* (pp. 169–178). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.18

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. *Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Odijk J, Piperidis S, editors. LREC 2022, 13th International Conference on Language Resources and Evaluation; 2022 June 20-25; Marseille, France. Paris: European Language Resources; 2022. 10 p.*

Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., Li, Y., & Chen, K. (2023). Rtmpose: Real-time multi-person pose estimation based on mmpose. https://doi.org/10.48550/ARXIV.2303.07399

Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., & Luo, P. (2020). Whole-body human pose estimation in the wild. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). Mediapipe: A framework for perceiving and processing reality. *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*. https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf

Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., & Chen, K. (2022). Rtmdet: An empirical study of designing real-time object detectors.

Mmpose: Openmmlab pose estimation toolbox and benchmark. (2020). https://github.com/open-mmlab/mmpose

Moryossef, A., Tsochantaridis, I., Dinn, J., Camgoz, N. C., Bowden, R., Jiang, T., Rios, A., Muller, M., & Ebling, S. (2021). Evaluating the immediate applicability of pose estimation for sign language recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3434–3440.

Núñez-Marcos, A., Perez-de-Viñaspre, O., & Labaka, G. (2023). A survey on sign language machine translation. *Expert Systems with Applications*, *213*, 118993.

Öqvist, Z., Riemer Kankkonen, N., & Mesch, J. (2020, May). STS-korpus: A sign language web corpus tool for teaching and public use. In E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Proceedings of the lrec2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives* (pp. 177–180). European Language Resources Association (ELRA). https://aclanthology.org/2020.signlang-1.29

Östling, R., Börstell, C., & Courtaux, S. (2018). Visual iconicity across sign languages: Large-scale automated video analysis of iconic articulators and locations. *Frontiers in psychology*, *9*, 331049.

How to talk the talk? A comparative overview of keyword usage in Hungarian and Slovenian parliamentary corpora

Abstract

This study conducts a comparative analysis of parliamentary speeches from Hungary and Slovenia using the ParlaMint corpus. By examining keywords and collocations in parliamentary discourse, the research aims to uncover the linguistic strategies employed by different political factions in both countries. The analysis reveals distinct linguistic patterns between the two countries and between coalition and opposition parties, highlighting the role of language in shaping political narratives. In Hungary, there is a notable divergence in the vocabulary used by coalition and opposition parties, with only a minimal overlap in keywords, suggesting a highly polarized political landscape. In contrast, Slovenian political discourse exhibits greater commonality between the two sides, though differences remain. The findings contribute to the understanding of political discourse in Central Europe, demonstrating how language both reflects and constructs political realities.

1 Introduction

Parliamentary speeches represent the cornerstone of democratic processes within parliamentary systems. They serve as a primary medium through which elected Members of Parliament (MPs) articulate their positions, debate legislation, and address the concerns of the populace they represent. These speeches are not only central in shaping legislative outcomes but also provide a rich dataset for linguistic and political analysis, reflecting the dynamic interplay between language, culture, and political discourse.

In recent years, the ParlaMint project (Erjavec et al., 2022, Erjavec et al., 2023) has emerged as a significant resource for scholars interested in the computational and linguistic analysis of parliamentary records. By offering a standardized, open-access corpus of parliamentary debates from various countries, ParlaMint facilitates comparative studies that were previously challenged by the lack of accessible and uniform data.

The ParlaMint corpora have initiated many studies across different disciplines. The study “Networks of Power: Gender Analysis in Selected European Parliaments,” (Skubic et al., 2022) examine the role of gender in parliamentary debates across three national parliaments (UK, Spain, and Slovenia) during specific terms. Kurtoğlu Eskişar and Çöltekin (2022) present initial findings from a quantitative analysis of emotions in the Turkish parliament over a decade (2011–2021). Kryvenko et al. (2023) outlines a research project focusing on political polarization within parliaments, using data from the ParlaMint 3.0 dataset and analyzing debates from Great Britain, Hungary, Ukraine, and Slovenia.

Our study uses the ParlaMint corpus to conduct a comparative analysis of parliamentary speeches from Hungary and Slovenia with regard to the consistency of political narrative. These countries, situated in Central Europe, provide intriguing cases for comparison due to their shared regional characteristics and distinct linguistic and political landscapes.

2 Corpora

The Hungarian National Assembly is the unicameral legislative body of Hungary, with 199 members, 12 advocates for nationalities and a chairman. The second version of the ParlaMint-HU corpus (included in ParlaMint 4.0) contains the minutes of the National Assembly from 2014 to 2023, comprising of terms 7, 8, and the first part of 9 of the Third Republic. The documents of the corpus thus range from 2014-05-06 to 2023-07-31, containing the official textual transcriptions of 514 sittings in this period.

The Slovenian Parliament, functioning as a bicameral legislature, is predominantly represented by the National Assembly, its principal legislative body. This assembly, comprising 90 deputies, including representatives for Italian and Hungarian ethnic groups, is tasked with legislative, electoral, and supervisory duties. It operates in electoral terms, holding both regular and extraordinary sessions to address various legislative agendas. Documentation of the National Assembly’s activities is comprehensive and publicly accessible. Annual reports detail its operations, structure, and legislative output, while the ParlaMint-SI corpus offers an extensive archive of debates from the 3rd to the 8th legislative sessions, including over 311,376 speeches and nearly 70 million words, available from October 27, 2000, to April 6, 2022 (Erjavec et al., 2023).

3 Methods

For the analysis we used NoSketch Engine, where we made two subcorpora in both ParlaMint-HU 4.0 and ParlaMint-SI 4.0 for coalition and opposition party speeches. We used data from the Hungarian corpus spanning the years 2022-23 and from the Slovenian corpus covering the period from March 2021 until May 2022.

The reasons for choosing two different periods were twofold: first, we wished to ensure that MPs’ roles are consistent (either ’coalition’ or ’opposition’), and second, we wanted to make subcorpora with nearly similar size. For the period we analyzed, in Slovenia, there were three coalition parties (SDS, NSi, Konkretno) and six opposition parties (DeSUS, LMŠ, SAB, SNS, Levica, SD). In Hungary, there were two coalition parties (Fidesz, KDNP) and six opposition ones (Jobbik, MSZP, DK, Mi Hazánk, LMP, Párbeszéd, Momentum).

The Slovene coalition subcorpus contains 1,353,893 tokens, the opposition subcorpus 2,505,118. Hungarian coalition subcorpus contains 1,336,078 tokens the opposition subcorpus 1,794,297.

In the analysis we used the Keywords and Collocations tools of NoSketch Engine. First, we compared opposition and coalition keyword lists, than we checked their collocations, that is words occurring in range of three positions to the left and right of the keywords and made some conclusions.

4 Results and discussion

Our main interest was to make an attempt to analyse certain features of political discourse, such as the proportion of common topics of the coalition and opposition parties, their language strategies and the consistency of their communication on a comparative basis, examining the situation in two countries. First, we compared 100 keywords, which we extracted from coalition and opposition subcorpora for both Slovenian and Hungarian speeches.

We compared the lists of keywords and found that in Slovenia, coalition and opposition parties shared 25 keywords, while in Hungary only 9, which leads to our first trivial conclusion that in Hungary, opposition and coalition parties have fewer topics in common than in Slovenia.

Assuming that this discrepancy is politically motivated and shows an attempt on performing control over the use of words (about controlling the narrative and the political power, see e.g. De Fina and Georgakopoulou, 2011), we checked whether keywords of both coalition and opposition speeches are new constructs in language. We found that comparing the above keyword lists of 100 with lists of lemmas extracted from corpora that were created more than a decade earlier resulted in some interesting findings. In the Slovenian keyword lists 24 lemmas [1] were new in the sense that they were not a part of Fidaplus corpus (Arhar et al., 2007). Nine of these lemmas represent proper names of institutions that simply did not exist in 2006, or of persons who were not active before 2006. Many of the lemmas are in close connection to the COVID pandemic. The remaining new words are more or less specialized terms like *gluhoslep*, *omrežnina* and *sosežigalnica*, and even a foreign word *fracking*, which stands for a specific oil mining method. The one and only new word, which is emotionally more expressive and might be politically motivated is *superagencija*, which is used only by opposition parties.

[1] (18 from oppositional: *LMŠ* ’Lista Marjana Šarca, a political party’, *covid* ’Covid’, *fracking* ’fracking’, *Dikaučič*, *Twitter*, *UKOM* ’Government Communication Office of the Republic of Slovenia’, *coviden* ’related to Covid’, *operandi* ’modus operandi’, *sosežigalnica* ’co-incineration plant’, *OCCAR* ’Organisation for Joint Armament Cooperation’, *gluhoslep* ’deaf-blind’, *PKP10* ’10th anti-Covid package’, *freking* ’fracking’, *pekape* ’PKP (anti-Covid package)’, *koronski* ’coronal (or related to covid)’, *PKP9* ’9th anti-Covid package’, *freaking* ’freaking’, *superagencija* ’super agency’ (with a 19th one, *dsovji*, which is a result of an erroneous lemmatization. It stands for *DSO* – ’retirement home’), 6 from coalitional speeches: *covid* ’Covid’, *protikoronski* ’anti-Covid’, *Dikaučič*, *UKOM*, *PKP10*, *omrežnina* ’network fee’.)

In contrast, the Hungarian keyword list contains 42 new words (21 for both coalition and opposition parties), meaning they are not part of the list of lemmas in the Hungarian Gigaword Corpus (Oravecz et al., 2014). A significant part of these keywords are proper nouns, yet there are many new compounds that are clearly politically motivated e.g.: *extraprofitadó* ('tax on extra profit'), *akkumulátorgyarmat* ('battery colony', which conveys the meaning of the country being treated as a colony by battery manufacturers), *rezsivédelem* ('overhead protection', which denotes a government policy aiming at keeping living costs low) or *dollárbaloldal* ('dollar left', suggesting that the opposition parties were given foreign financial support). As seen from the examples, word-for-word translation of these compounds makes little sense in English, therefore a longer explanation is necessary to clarify their meaning. The intention behind coining new terms by glueing together words is to suggest that the denoted notions are strongly connected. As if the opposition politicians would all be financed from abroad and the manufactoring of batteries would necessarily entail that the country is in inferior position to the investors.

If we take a closer look at the 10 most frequent keywords of both sides, we see that in Hungary there is only one keyword shared (*energiaválság* - 'energy crisis'), whereas in Slovenia there are 2 (*epidemija* 'pandemic' and *covid*). This is not surprising, as 2021 and partly 2022 were marked by the negative consequences of the pandemic. Sharing only one keyword does not mean that the sides talked about different topic. In Hungary, many times the topics were the same, but the words were different. E.g. with regard to costs the opposition's keywords were *rezsiemelés*, *rezsinövelés* 'overhead increase', whereas coalition MP's used *rezsivédelem* 'overhead protection', *rezsicsökkentés* 'overhead decrease', *rezsitámogatás* 'overhead support'. The two sides therefore talk about the same topic, yet, with different words.

The conscious use of words is apparent in the list of collocations of 'energy crisis' as used by coalition and opposition parties. Surprisingly, there is not a single common collocation (among the 10 most frequent ones) between the two lists, which shows that the issue was discussed using different narratives.

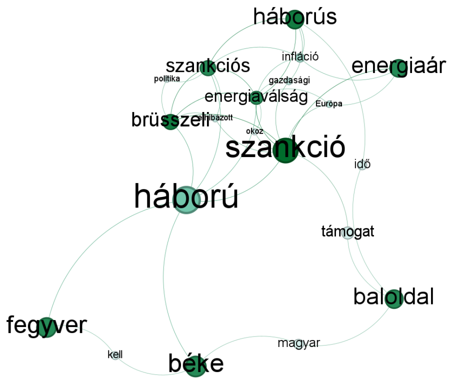
- Collocations of *energiaválság* 'energy crisis', opposition: *klímaválság* 'climate crisis', *közép* 'middle / center', *kellős* 'very', *válság* 'crisis', *beszél* 'talk', *lesz* 'will be', *idő* 'time', *Magyarország* 'Hungary', *megélhetési* 'living', *jelenlegi* 'present'
- Collocations of *energiaválság* 'energy crisis', coalition: *háború* 'war', *szankció* 'sanction', *Európa* 'Europe', *infláció* 'inflation', *okoz* 'causes', *gazdasági* 'economic', *szankciós* 'sanction (adjective)', *háborús* 'war (adjective)', *miatt* 'because of', *brüsszeli* 'Brussels (adjective)'

Figure 1a and 1b depict the graphical representation of the previously analyzed keywords and their collocations. The nodes represent the keywords and their collocations, while the edges connect the keywords to their collocations. For instance, in the opposition graph of the Hungarian dataset, there are five edges leading to *Magyarország* ('Hungary') because it is listed among the top ten collocations for five opposition keywords. The edges were not weighted according to the frequency of the collocations. The size of a label or node is proportional to its betweenness centrality value, indicating the node's importance (the larger, the more significant). The color depth is proportional to the Degree metric, with lighter colors indicating fewer connections.

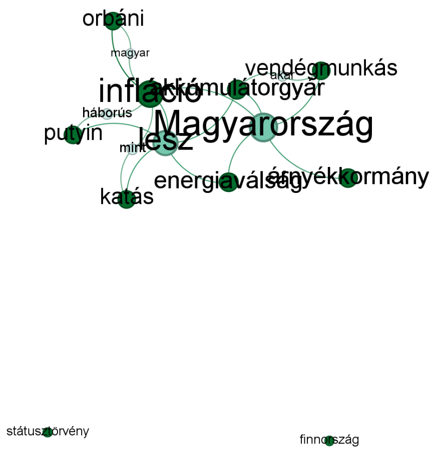
If we measure the density of the graphs, we see, that the graph for the Hungarian coalition has the highest value (0.038) and the Hungarian opposition the lowest (0.021). In case of Slovenia, coalition parties turn out to have a more focused communication (0.022) that opposition ones (0.032), too, yet the difference between them is smaller.

5 Conclusion

This study presented a comparative analysis of parliamentary speeches from Hungary and Slovenia, using the ParlaMint corpus and highlights remarkable differences in the political discourse between the two countries, with a notable divergence in the vocabulary used by different political factions within Hungary, indicating a highly polarized political landscape. In contrast, Slovenian political discourse exhibits a higher degree of commonality between coalition and opposition parties.



(a) Graph of the coalition's keywords and collocations. As can be seen, *háború* 'war' and *szankciós* 'sanction, adjective' rule the graph with the most importance and the most connections. Each node can be reached from any other node –representing the focused communication style of the governing coalition.



(b) Graph of the opposition's keywords and collocations. There appear to be three disjunct subgraphs of words, indicating that the opposition's speeches focus on three different topics, in contrast with the coalition's focused and more strictly defined communication lines.

References

Arhar, Š., Gorjanc, V., & Krek, S. (2007). FidaPLUS corpus of Slovenian: the new generation of the Slovenian reference corpus: its design and tools.

De Fina, A., & Georgakopoulou, A. (2011). Narrative power, authority and ownership. In *Analyzing Narrative: Discourse and Sociolinguistic Perspectives* (pp. 125–154). Cambridge University Press.

Erjavec, T., Kopp, M., Ogrodniczuk, M., & Osenova, D., Petya ... Fišer. (2023). Multilingual comparable corpora of parliamentary debates ParlaMint 4.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1859

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Matyáš ... Marx, & Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings. *Lang. Resour. Eval.*, *57*(1), 415–448. https://doi.org/10.1007/s10579-021-09574-0

Kryvenko, A., Evkoski, B., Bordon, D., & Meden, K. (2023). Splitting lips: polarization through parliamentary speech: [plakat na mednarodni konferenci "Helsinki Digital Humanities Hackathon #DHH23", Helsinki, Finska, 24. 5.-2. 6. 2023] [Nasl. z nasl. zaslona]. https://www.helsinki.fi/assets/drupal/2023-06/dhh23-parliament-poster.pdf

Kurtoğlu Eskişar, G. M., & Çöltekin, Ç. (2022, June). Emotions Running High? A Synopsis of the state of Turkish Politics through the ParlaMint Corpus. In D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the workshop parlaclarin iii within the 13th language resources and evaluation conference* (pp. 61–70). European Language Resources Association. https://aclanthology.org/2022.parlaclarin-1.10

Oravecz, C., Váradi, T., & Sass, B. (2014). The Hungarian Gigaword Corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1719–1723. http://www.lrec-conf.org/proceedings/lrec2014/pdf/681%5C_Paper.pdf

Skubic, J., Angermeier, J., Evkoski, B., Bruncrona, A., & Leiminger, L. (2022). Networks of Power - Gender Analysis in Selected European Parliaments. *2nd Workshop on Computational Linguistics for Political Text Analysis (CPSS-2022).*

Unlocking the Corpus: Enriching Metadata with State-of-the-Art NLP Methodology and Linked Data

Jennifer Ecker[1], **Stefan Fischer**[2], **Pia Schwarz**[1], **Thorsten Trippel**[1], **Antonina Werthmann**[1], **Rebecca Wilm**[1]
[1]Leibniz Institute for the German Language (IDS), Mannheim, Germany
[2]Saarland University, Saarbrücken, Germany
{ecker,schwarz,trippel,werthmann,wilm}@ids-mannheim.de
stefan.fischer@uni-saarland.de

1 Introduction

In research data management, descriptive metadata are indispensable to describing data and are a key element in preparing data according to the FAIR principles (Wilkinson et al., 2016). Extracting *semantic* metadata from textual research data is currently not part of most metadata workflows, even more so if a research data set can be subdivided into smaller parts, such as a newspaper corpus containing multiple newspaper articles. Our approach is to add semantic metadata at the text level to facilitate the search over data. We show how to enrich metadata with three NLP methods: named entity recognition, keyword extraction, and topic modeling. The goal is to make it possible to search for texts that are about certain topics or described by certain keywords, or to identify people, places, and organisations mentioned in texts without actually having to read them and at the same time facilitate the creation of task-tailored subcorpora. To enhance this usability of the data we explore options based on the German Reference Corpus *DeReKo*, the largest linguistically motivated collection of German language material (Kupietz & Keibel, 2009; Kupietz et al., 2010, 2018), which contains multiple newspapers, books, transcriptions, etc., and enrich its metadata on the level of subportions, i.e. newspaper articles. We received access to a number of data files in DeReKo's native XML format, I5. To develop the methodology, we focus on a single XML file containing all issues of one newspaper of a whole year.

The following sections only give an overview of our approach, we intend, however, to provide a detailed description of the experiments and the selection of data in a subsequent longer contribution.

2 Motivation

Specific research questions may focus on parts of a corpus. However, there is no general criterion for substructuring a corpus, as this is highly dependent on the research questions. For someone interested in the style of specific authors, the substructuring of such a corpus would be best if all articles or contributions would be clustered by their author; for someone interested in specific topics, the clustering should be by topic, for specific time frames by dates, etc. The internal structure of the original corpus data, however, also allows for other partitions, such as identifying all newspaper articles published on a specific date, belonging to specific sections, etc. These structures can be identified by their internal unique text *sigles*, which are part of the XML representations of the data. Hence, a sigle is a unique identifier for a subpart, either at a corpus, document, or text level, the latter corresponding to individual newspaper articles, for example. Our goal is to enrich the corpus semantically on the text level in order to allow researchers to create subcorpora tailored precisely to their scientific needs.

3 Approach

As a starting point to enrich DeReKo with semantic metadata, we focus on extracting topics, keywords, and three types of named entities: academic institutions, research areas, and persons with an academic background. We believe that these might be useful entry points for researchers to partition the reference corpus to fit their particular research questions. Additionally, applying three different NLP methods allows us to explore how we can implement a metadata extension that captures semantic metadata besides the existing catalog metadata. This extension also includes the possibility to record potential links between the extracted semantic information and other external data sets such as Wikidata (Vrandečić & Krötzsch, 2014), the Integrated Authority File (Gemeinsame Normdatei; GND; Behrens-Neumann & Pfeifer, 2011), the Research Organization Registry (ROR; Lammey, 2020), the Open Researcher and Contributor ID (ORCID; Haak et al., 2012) or even to lexical units from GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010), a lexical-semantic network for German. To link extracted topics, keywords, and named entities, the respective identifiers (QIDs, GND IDs, IDs from ROR and ORCID, GermaNet IDs) of corresponding items from all these external knowledge bases can be encoded per text sigle in semantic stand-off annotation files.

To find an appropriate strategy to enrich the metadata with such semantic information, we explored four main approaches: (i) Integrating semantic information directly into the metadata header of the I5 files; (ii) Storing enriched metadata in separate metadata files (in formats such as CMDI or JSON-LD) for each individual sigle of the I5 file; (iii) Automatically generating CMDI files with descriptive metadata extracted from the I5 files, followed by expanding them with semantic information; and (iv) Generating semantic information through real-time analysis. Option (iii) was chosen: The rationale for this approach is that it allows for the storage of both types of metadata (descriptive and semantic) in the same file without altering the primary I5 files; changing the CMDI schema is possible at any time, and files can be converted to alternative formats, such as JSON-LD or HTML. CMDI files also facilitate sharing metadata under open licenses, promoting accessibility and transparency while respecting legal restrictions on the original data. Moreover, the chosen option avoids the significant technical and result-related challenges of high computational resource requirements and infrastructure costs to generate the information in real time. Although enriching semantic CMDI metadata requires intricate data modeling and interpretation, it can be effectively implemented.

4 Data

For our experiments we selected a data sample from DeReKo, the 2020 volume of the newspaper corpus Mannheimer Morgen (M20), published under the QAO-NC license (Kupietz & Lüngen, 2014) allowing for query-and-analysis-only, academic and non-commercial use. According to DeReKo's structure, M20 is a single I5-formatted XML file containing several individual newspaper articles each identified by their text sigle, e.g. M20/APR.00192, which consists of the corpus identifier M20, the document identifier, corresponding to the month in which the article was released, and a five-digit text identifier. In total the M20 subcorpus comprises 44,383 texts.

5 Experiments: Topic Modeling, Keyword Extraction, and Academic NER

We employ topic modeling to group the articles into categories and use the topic model tool Top2Vec (Angelov, 2020) to assign topics to the articles. Top2Vec divides the articles into 348 topics and assigns the cosine similarity to every article. After a manual inspection, we use hierarchical topic reduction to reduce the topics from 348 to 150 to circumvent that about half the topics are semantically too close to each other. The output of the topic model is a number of unnamed topics and corresponding topic words for each topic. We then prompt a large language model to do the labeling. The used model is based on Meta's Llama 2 model and twenty topic words form the basis for computing the label. The topic label can be one of the topic words or a word derived from them. If the generated label is off topic (e.g. 'unleashing' for topic words about tracing/describing a person/an offender), we define a label manually.

To extract keywords, up to ten uni- or bigram keywords are extracted for each article by combining YAKE! (Campos et al., 2018a, 2018b, 2020), a state-of-the-art unsupervised keyword extraction method that assigns a score to each possible keyword based on statistical features, with a filter based on spaCy (Honnibal et al., 2020) part-of-speech tags to exclude any parts of speech other than nouns and proper nouns. In order to avoid inflected forms such as 'Bundesfinanzministeriums' ('Federal Ministry of Finance's'), the resulting keywords are lemmatised using spaCy's lemmatizer.

In order to recognize academic named entities, we fine-tune a German BERT$_{\text{BASE}}$ model as, to our knowledge, no suitable German NER model exists that recognizes the three entity types we require. To obtain training data, sentences are filtered out of 10,000 randomly selected texts from DeReKo with the help of an off-the-shelf NER model from the Stanza NLP package (Qi et al., 2020) and word lists containing prototypical mentions for each of the entity types academic person (PER-RES), academic institution (ORG-RES), and research area (AREA-RES) to detect candidate entities. During post-processing, candidate entities are manually reviewed, resulting in a data set of 4,928 sentences with a total of 7,199 tags. Using the spaCy transformer library, we fine-tune the model *de_dep_news_trf* and perform the evaluation on the 489 sentences of the test split, yielding an overall micro-averaged F1 score of 92.45%.

6 Results and Outlook

Enriched metadata is crucial for enhancing corpus usability, accessibility and value. Users can easily access comprehensive information on a corpus, including topics, relevant keywords, and named entities, without the need to read each text individually. Below are our results and an outlook on the next steps.

For the method of topic labeling, the majority of the generated names for the topics are suitable. Nevertheless, we changed the label in 39 out of 150 cases (26%). To improve the topic words, one option is to apply lemmatisation before or after topic modeling.

Due to the subjective and impractical nature of manually determining a gold standard for keywords for a large amount of texts, no qualitative evaluation is conducted regarding keyword extraction. However, an examination of model-generated example keywords suggests they generally serve their purpose by indicating article topics. Yet, they may contain errors in part-of-speech tagging and lemmatisation, and the suitability of the chosen YAKE! parameters remains uncertain.

When applying the fine-tuned NER model to the M20 subcorpus, at least one academic named entity is tagged in almost 40% of the 44,383 newspaper articles. Most of the tags, over 20,000, fall upon the type PER-RES, almost 10,000 items are tagged with ORG-RES, and a bit more than 3,000 with the entity type AREA-RES. Due to the size of M20, no qualitative analysis was made. To find error patterns in the results, a more detailed analysis is required, also including the question of consequences of adding more training data. An indication of the model's confidence could be helpful for assessing the quality of assigned tags.

Depending on the method, processing of the M20 subcorpus took between 6 hours and 3 days, which could be further optimized regarding runtime and memory usage. Scaling this for the entire DeReKo corpus would require optimization, possibly through parallel computing. Similar challenges exist for temporary files generated during preprocessing I5 files before running the NLP processes.

Our experiments demonstrate the feasibility of extracting metadata from large corpora, with potential future use in corpus analysis tools for identifying subcorpora. This methodology is applicable to various national and large corpora, including those with legal constraints, like DeReKo. The method, adaptable across languages and corpora with similar structures, awaits scaling for broader use. Integration with linked data sources remains open. This includes plans to incorporate GermaNet, ontologies, and authority files for enhanced metadata connectivity. Scaling experiments to the full dataset requires stable, automated processes, building on successful initial tests, which still have to be performed. Future investigations will explore applying these methods to different corpus types, including endangered language corpora such as DOBES, leveraging multi-tier annotations for metadata enrichment.

References

Angelov, D. (2020). Top2Vec: Distributed representations of topics. https://doi.org/10.48550/arXiv.2008.09470

Behrens-Neumann, R., & Pfeifer, B. (2011). Die Gemeinsame Normdatei – ein Kooperationsprojek (D. Nationalbibliothek, Ed.). *Dialog mit Bibliotheken*, *23*(1), 37–40.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). Yake! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257–289. https://doi.org/10.1016/j.ins.2019.09.013

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018a). A text feature based automatic keyword extraction method for single documents. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 684–691). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_63

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018b). Yake! Collection-independent automatic keyword extractor. In G. Pasi, B. Piwowarski, L. Azzopardi, & A. Hanbury (Eds.), *Advances in information retrieval* (pp. 806–810). Springer International Publishing. https://doi.org/10.1007/978-3-319-76941-7_80

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264.

Hamp, B., & Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 9–15. https://aclanthology.org/W97-0802

Henrich, V., & Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, 2228–2235.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)* (pp. 1848–1854). European Language Resources Association (ELRA) 2010.

Kupietz, M., & Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi (Ed.), *Workings Papers in Corpus-based Linguistics and Language Education* (pp. 53–59, Vol. 3). Tokyo University of Foreign Studies 2009.

Kupietz, M., & Lüngen, H. (2014). Recent developments in DeReKo. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2378–2385.

Kupietz, M., Lüngen, H., Kamocki, P., & Witt, A. (2018). The German Reference Corpus DeReKo: New Developments -– New Opportunities. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 4353–4360.

Lammey, R. (2020). Solutions for identification problems: A look at the research organization registry. *Science Editing*, *7*(1), 65–69. https://doi.org/10.6087/kcse.192

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*.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85. https://doi.org/10.1145/2629489

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, *3*.

Choosing the Right Tool for You: Informed Evaluation of Text Analysis Tools

Angel Daza
Netherlands eScience Center
j.daza@esciencecenter.nl

Antske Fokkens
Vrije Universiteit Amsterdam
antske.fokkens@vu.nl

Abstract

Natural Language Processing (NLP) showcases many promising tools and methods for text analysis. Scholars from diverse fields want to use NLP to help with their research and are confronted with a wide availability of ready-to-use models that claim excellent performance in standard benchmarks. Consequently, choosing an appropriate tool has become a task on its own. Our main goal is to exemplify a methodology that stimulates critical evaluation and detailed analysis of automatic outputs of NLP tools. Particularly, we analyze the case of choosing the best Named Entity Recognizer (NER) for a corpus of biographies written in Dutch. Our use case is an example of how to make informed decisions by considering different aspects of custom datasets at the instance and aggregated levels; improving the outcomes of the original research question.

1 Introduction

Recent developments in NLP have greatly increased one of CLARIN's primary goals of providing (relatively) easy-to-use language technology. This has led to an increase of such technologies being used in various domains, such as Digital Humanities (DH) (Colavizza et al., 2021; Ehrmann et al., 2023; Schweter et al., 2022). However, the same rapid advancement of NLP has left the area with a weak spot regarding detailed and careful evaluation. Standard benchmarks and metrics often do not provide sufficient insight for users to establish which tool works best for their specific use case, nor whether this tool performs well enough for a methodological set-up at all (Fokkens et al., 2014). In our view, supporting users in establishing what tool is most suitable for their use case forms an essential part of CLARIN's mission of making tools available to researchers. In this abstract, we provide an approach that fosters critical analysis of automatic outputs generated by NLP tools at both an instance-level and aggregated level. This allows users to zoom into the strengths and weaknesses of different models, enabling well-informed decisions regarding the suitability of these models for actual use cases. For the sake of clarity, we put ourselves in the place of a historian aiming to automatically build networks of people (`PER`), organizations (`ORG`), and places (`LOC`) using digital biographies; however, many of the ideas exposed here can be generalized to most token classification and span-based detection tasks in NLP.

2 Methodology

2.1 Case Study: NER for Dutch Biographies

The Biography Portal of the Netherlands[1] (BPN) is a digital corpus of thousands of biographies written in Dutch between the 18th and the 21st centuries.It comes with a set of typical DH challenges: Language variety and dynamic changes through the centuries, a mixture of record typologies (a high diversity in style, old and modern Dutch), significant divergence in biography length, many (idiosyncratic) abbreviations, and rare entities that do not necessarily exist nowadays, etc. To validate our research, we also have access to a human-labeled dataset based on a subset of 346 biographies of various lengths, this subset was generated by stratified sampling to keep the original dataset's source distribution, ensuring we have

[1] http://www.biografischportaal.nl

examples from each collection. Table 1 provides an overview of the data where we will run all further analyses.

Category	Count	Mean	Median	Max	StdDev
PER	5,743	17	10	92	18.7
LOC	3,879	11	8	77	12.0
ORG	2,196	6	2	58	9.8
ALL	11,818	34	21	164	35.8
Tokens	189,507	548	245	3,126	613.8
Sents	8,210	23	21	210	13.5
Docs	346	-	-	-	-

Table 1: General statistics of our manually annotated corpus. We include the average, maximum, and median of occurrences per document to illustrate the heterogeneity of the dataset at the document level.

2.2 NER Models

We focus on `PER`, `LOC`, and `ORG` because their definition is less controversial than other entity categories. We consider four different models that deliver (at least) the three desired NER labels. All models are readily available to use out-of-the-box, making them very attractive for non-NLP researchers. Commonly, the available NER taggers report scores in publicly available benchmarks for NER, such as the conll-02 (Tjong Kim Sang, 2002); however, that does not necessarily say anything about the performance that they will have on our specific use case. To clarify that, we will use *Flair NLP* (Akbik et al., 2019) *Stanza-NER* (Nothman et al., 2013), *Fine-tuned XLM-R* for NER[2] and we prompt *gpt3.5-turbo* (Brown et al., 2020) for obtaining NER labels on our dataset and compare their outputs.

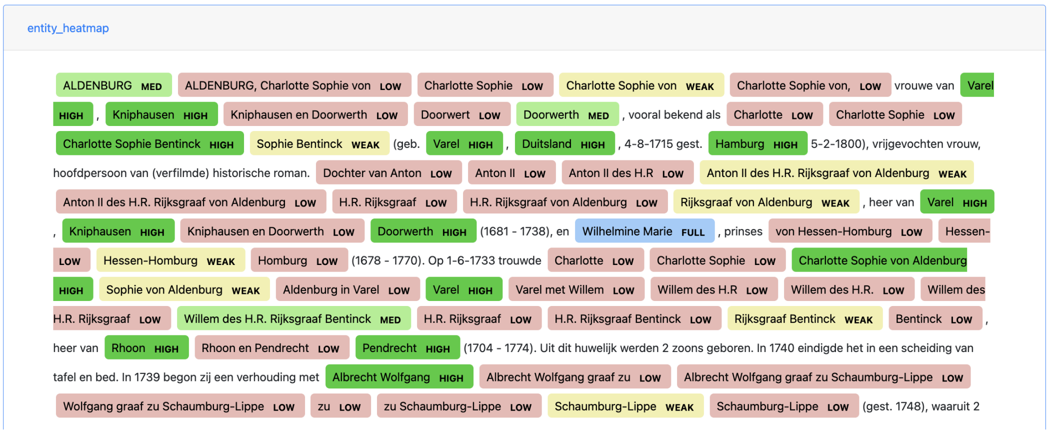
2.3 Inspection of Model Behavior

We compare the outputs in parallel to spot interesting behavior. An instance can be a sentence, paragraph, or document. By looking closely at relevant instances, we can investigate what errors or tagging biases the different models exhibit and make decisions accordingly. It would be impossible to inspect every single instance closely; therefore, we obtain the most interesting instances to analyze by computing the divergence of all model predictions in each instance: we predict that the more the models disagree, the more interesting the case will be. This is possible even when no human-annotated data is available.

Divergence Matrix. We have a collection of P instances $\{p_0, p_1, \ldots, p_P\}$, a set of N models denoted as $\{m_0, m_1, \ldots, m_N\}$, and M evaluation metrics $\{m_1, m_2, \ldots, m_M\}$. We generate a matrix Z, with dimensions $N \times (P * M)$. In this context, a row is an instance p_i, and each column $Z_{i,j,k}$ signifies the performance score of model m_j when evaluated using metric m_k on instance p_i with $i = 0, 1, \ldots, P$; $j = 0, 1, \ldots, N$; $k = 0, 1, \ldots, M$. Consequently, the matrix encapsulates the evaluated models' performance across all instances and metrics. The three metrics computed are **Entity Frequency:** we get the raw counts of Named Entities found in each document; **Entity Density:** the entity frequency divided by the number of tokens in the instance to get a weighted metric; and **Entity Divergence:** is the standard deviation of the frequency arrays from each $model_n$. The reasoning is that models will obtain the same amount of labels for easy instances (all models will agree), whereas instances with complex cases or cases with noise will have a high divergence.

Entity Heatmap. We classify text spans (entity candidates) into five buckets according to the certainty of correctness using models in a voting mechanism: the entities that have the votes of all N models are highlighted in blue (FULL), in dark green are the entities that $N - 1$ models identified (HIGH), in light green $N - 2$ (MED), in yellow $N - 3$ (WEAK) and the rest of entities are labeled in red (LOW) indicating low certainty based on only one or two models identifying them. This way, we can immediately visualize which spans are problematic in the actual document context where they were detected.

[2] https://huggingface.co/Davlan/xlm-roberta-base-ner-hrl



entity_heatmap

ALDENBURG MED ALDENBURG, Charlotte Sophie von LOW Charlotte Sophie LOW Charlotte Sophie von WEAK Charlotte Sophie von, LOW vrouwe van Varel
HIGH , Kniphausen HIGH Kniphausen en Doorwerth LOW Doorwert LOW Doorwerth MED , vooral bekend als Charlotte LOW Charlotte Sophie LOW
Charlotte Sophie Bentinck HIGH Sophie Bentinck WEAK (geb. Varel HIGH , Duitsland HIGH , 4-8-1715 gest. Hamburg HIGH 5-2-1800), vrijgevochten vrouw,
hoofdpersoon van (verfilmde) historische roman. Dochter van Anton LOW Anton II LOW Anton II des H.R LOW Anton II des H.R. Rijksgraaf von Aldenburg WEAK
Anton II des H.R. Rijksgraaf von Aldenburg LOW H.R. Rijksgraaf LOW H.R. Rijksgraaf von Aldenburg LOW Rijksgraaf von Aldenburg WEAK , heer van Varel HIGH
, Kniphausen HIGH Kniphausen en Doorwerth LOW Doorwerth HIGH (1681 - 1738), en Wilhelmine Marie FULL , prinses von Hessen-Homburg LOW Hessen-
LOW Hessen-Homburg WEAK Homburg LOW (1678 - 1770). Op 1-6-1733 trouwde Charlotte LOW Charlotte Sophie LOW Charlotte Sophie von Aldenburg
HIGH Sophie von Aldenburg WEAK Aldenburg in Varel LOW Varel HIGH Varel met Willem LOW Willem des H.R LOW Willem des H.R. LOW Willem des
H.R. Rijksgraaf LOW Willem des H.R. Rijksgraaf Bentinck MED H.R. Rijksgraaf LOW H.R. Rijksgraaf Bentinck LOW Rijksgraaf Bentinck WEAK Bentinck LOW ,
heer van Rhoon HIGH Rhoon en Pendrecht LOW Pendrecht HIGH (1704 - 1774). Uit dit huwelijk werden 2 zoons geboren. In 1740 eindigde het in een scheiding van
tafel en bed. In 1739 begon zij een verhouding met Albrecht Wolfgang HIGH Albrecht Wolfgang graaf zu LOW Albrecht Wolfgang graaf zu Schaumburg-Lippe LOW
Wolfgang graaf zu Schaumburg-Lippe LOW zu LOW zu Schaumburg-Lippe LOW Schaumburg-Lippe WEAK Schaumburg-Lippe LOW (gest. 1748), waaruit 2

Figure 1: We can visualize which entities most models prefer and which cause more disagreement among them.

Parallel Span Comparison. Another source of confusion is the label of the recognized spans. Take the span `Varel, Kniphausen en Doorenwerth` in Figure 1, which is basically an enumeration of places that Charlotte von Aldenburg is a *lady of*. Some models interpreted that this segment could mean she is the *wife of* and thus classify these Entities as people None of the NER systems are trained to deal with these cases, we thus see a higher disagreement in how they are labeled and, hence, the high divergence across PER and LOC in this document. Inspecting such examples already provides hints to fix mistakes, such as filtering in a post-processing step, for example.

3 Findings and Discussion

We use a set of visualizations built on top of everyday tools to inspect the instances and make the process more user-friendly[3]. We use spreadsheets to show matrices. Additionally, we created a small Flask web app that integrates visualizations from `displaCy`[4] to see the spans on the text and Google Charts[5] to show some global statistics. These visualizations can help identify where the source of divergences can be, and remind us that we should check for the following aspects when evaluating. Depending on which aspects are more relevant to our use case, different systems can come up as the best solution for our specific use case:

3.1 Tokenization Matters

Since the CoNLL-02 (Tjong Kim Sang, 2002) shared task, NER has been approached in NLP as a sequence labeling task, where each token is assigned one label in the IOB format (or related) (Ramshaw & Marcus, 1995). Notably, the most common scenario for non-NLP users is to apply an out-of-the-box tool to raw text. Our experiments show that evaluation on raw text gives significantly lower scores than the evaluation assuming tokenization as a given. We consider the *Span match* mode to be closer to what external users of tools need.

3.2 Full Match vs Partial Match

Only entities that match entirely the gold span label are considered correct in a full match setting. This strict match policy can be too harsh for cases where the classifier almost got the whole entity with the correct label but missed a couple of tokens compared to the gold (which can also be fixed with a simple post-processing rule). For example, if a label `Charlotte Sophie` instead of `Charlotte Sophie von Aldenburg` as `PER` in her own biography can easily be mapped to the correct person leading to a fully correct outcome for making networks, thus partial match evaluation could be enough for this case.

[3] The code for tools that support this analyses is available at https://github.com/angel-daza/bios-dutch
[4] https://demos.explosion.ai/displacy-ent
[5] https://developers.google.com/chart/

3.3 Bag of Entities

In some cases, e.g. identifying relevant documents or creating networks based on loose connections, we only need to encounter the correct entity in the text once. It then does not matter if *Amsterdam* occurs many times in the document, we only need to recognize it once as a LOC to draw an edge between this person and *Amsterdam*. Here, evaluating without the need of validating spans can be enough.

3.4 Precision vs Recall

A related aspect is whether high precision (not many false positives) or high recall (not many items missed, or few false negatives) is more important. Because the NER task is span identification and span labeling, there are different causes of error: i) The span was correctly identified, but the label is wrong, ii) The span is wrongly identified, but the label is correct, iii) The span was wrongly identified and the label is wrong, iv) The labeler did not tag at all the entity. If this information remains packed in a single P, R, or F1 score, we lose the ability to analyze the errors. Visualizing the FPs and FNs of individual documents separately can show what kind of mistakes are made so we can act accordingly to fix them.

4 Conclusion

In this abstract, we call for a more detailed evaluation of NLP span classification tasks. We apply several out-of-the-box NER models to a corpus of Dutch Biographies and compare different options for evaluation. We aim to illustrate the importance of inspecting the output of models at various levels when investigating whether their output provides the information that is needed with sufficient reliability. We also aim to show that a higher F1 score in a benchmark does not necessarily mean that the model will also be the best choice for our specific use case. We shared the code where we performed all these analyses. The paper's full version will further support these claims with the full results of our experiments and illustrative examples, which we currently left out due to lack of space.

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.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and ai: An overview of current debates and future perspectives. *J. Comput. Cult. Herit.*, *15*(1).

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, *56*(2).

Fokkens, A., Ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., et al. (2014). Biographynet: Methodological issues when nlp supports historical research. *LREC*, 3728–3735.

Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artif. Intell.*, *194*, 151–175.

Ramshaw, L., & Marcus, M. (1995). Text chunking using transformation-based learning. *Third Workshop on Very Large Corpora*. https://aclanthology.org/W95-0107

Schweter, S., März, L., Schmid, K., & Çano, E. (2022). Hmbert: Historical multilingual language models for named entity recognition.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. https://aclanthology.org/W02-2024

Improving Phrase Structure Parsing for Icelandic

Ingunn Jóhanna Kristjánsdóttir
University of Iceland
ijk4@hi.is

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Anton Karl Ingason
University of Iceland
antoni@hi.is

Abstract

We present an improved state-of-the-art in phrase structure parsing for Icelandic, building on the Icelandic Parsed Historical Corpus (IcePaHC), a CLARIN resource, as well as previous milestones presented at the CLARIN Annual Conference in the past. The present parsing system utilizes the Stanford Stanza system as well as IceBERT, a freely available Icelandic BERT Model. We describe previous work, the different configurations used for the present work, as well as the setup that yielded the best outcome, an F1 score of 90.38%.

1 Introduction

We present a new phrase structure parser for Icelandic, based on the Stanford Stanza system (Qi et al., 2020), which achieves a new state-of-the-art performance for the task in this language.

When considering resources such as syntactic parsers, Icelandic is a language with a relatively few speakers and it has lagged behind other bigger language communities in terms of developing crucial infrastructure. At the turn of the century, Icelandic Language Technology was virtually non-exisitent (Loftsson et al., 2009; Rögnvaldsson, 2010), at a time when English had resources like the Penn treebank (Marcus et al., 1993) and associated software innovations. However, around 2010, limited support for basic resources such as taggers (Loftsson, 2008) and lemmatizers (Ingason et al., 2008) had emerged for Icelandic and in 2011 a manually corrected constituency treebank, the Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011), was released, leading the way for further development of parsing resources, and since then substantial progress has been made, thanks to a systematic Language Technology Programme (Nikulásdóttir et al., 2022), supported by Icelandic Government, an effort which has led to a large number of new CLARIN resources. While Icelandic still lags behind the major languages (Rehm & Way, 2023), this most recent period, along with international developments in new techniques, has involved fruitful development of new parsing tools.

In this paper, we present the Icelandic Stanza Phrase Structure Parser, an experiment where we train Stanza on the IcePaHC treebank and reach an F1 score of 90.38%, an improvement over earlier experiments. The paper is organized as follows. In Section 2, we review some background on Icelandic parsing resources and previous parsers for Icelandic. In Section 3, we describe our evaluation of the system and how it compares to other systems. In Section 4, we report on the findings. Finally, Section 5 concludes.

2 Background

2.1 Icelandic parsing resources

The Penn Treebank (Marcus et al., 1993) was the most significant resource for starting the research program that develops constituency parsers and it remains a key test case for new parsers to this date. Yet, some of the same experts who developed this treebank moved on to an annotation scheme that improves on some of design decisions in the Penn Historical Corpora (Kroch & Taylor, 2000; Kroch et al., 2004). This includes a relatively more flat structure that abstracts away from ambiguities such as

PP-attachment in many cases. This latter family of treebanks became the basis for the annotation scheme of the Icelandic Parsed Historical Corpus (IcePaHC), released in 2011 (Rögnvaldsson et al., 2011, 2012; Wallenberg et al., 2011). While we focus here on constituency treebanks, it is worth noting that datasets annotated with the Universal Dependencies annotation scheme became available later. One such treebank is described in Jónsdóttir and Ingason (2020), and another one, created by converting IcePaHC to the UD format, is described in Arnardóttir et al. (2020).

2.2 Previous parsers for Icelandic

While most parsers are data-driven and trained on corpora, the first solution for Icelandic was a rule-based shallow parser, IceParser (Loftsson & Rögnvaldsson, 2007). The first attempts at full phrase structure parsing involved a combination of Icelandic and Faroese data in Ingason et al. (2014), in which a section of data from IcePaHC was combined with the Faroese Parsed Historical Corpus (FarPaHC) (Sigurðsson et al., 2012). However, the first parsing experiment that used the full dataset of IcePaHC was described in Jökulsdóttir et al. (2019) using the Berkeley Parser as the central piece of a parsing pipeline. In the following year, an improved version of this system was created, based on the Berkeley Neural Parser (Arnardóttir & Ingason, 2020). This system reached an 84.74% F1 score, aided by a multilingual Bert model, with a recall of 84.43% and a precision of 85.07%. Without using the word embeddings, the system reached an F1 score of 82.18%.

In the realm of rule-based systems, Þorsteinsson et al. (2019) introduced a parser based on a wide-coverage context-free grammar for Icelandic. Furthermore, parsers that focus on dependency parsing have been released in recent years, building on the conversion of IcePaHC to UD. This includes the experiment in Arnardóttir et al. (2023) as well as the Stanford Stanza UD Parser (Qi et al., 2020).

As having trained word embeddings for a particular language is useful for improving performance, it has been a limiting factor in earlier parsing experiments that a contextual word embedding model for Icelandic was not available. Therefore, it is important for current work on Icelandic parsing that an Icelandic BERT model has now been made available, named IceBERT (Snæbjarnarson et al., 2022). We make use of IceBERT in our setup. Snæbjarnarson et al. (2022) also carried out a constituency parsing experiment on the GreynirCorpus test set and reached an F1 score of 90.02%. Their finding is similar to ours but not directly comparable because of the different nature of the GreynirCorpus test set.

3 Evaluation

Before training the Stanza parser on the Icelandic phrase structure data in IcePaHC, we had to make sure that the data was in an appropriate format. This meant that we had to clean up any empty nodes in the trees, as the parsing task is separate from recovering the identity of silent elements and traces of syntactic movement. We furthermore split the IcePaHC treebank into a training set, development set and test set. IcePaHC consists of one million words in 73,012 matrix clauses and 80% of these clauses are used for the training set, 10% for the development set and 10% for the test set. IcePaHC consists of data from different centuries (dated 1150–2008) and to guarantee an even distribution in the three sets, every tenth part of the corpus is divided between them. All computations were performed on resources provided by the Division of Information Technology of the University of Iceland through the Icelandic Research e-Infrastructure project, funded by the Icelandic Centre of Research.

4 Results

The results of the evaluation are shown in Table 1, contrasted with the earlier experiment of Arnardóttir and Ingason (2020). The table indicates, for each experimental setup, whether the IceBERT model was used (and whether it was the basic IceBERT model or the larger IceBERT Large), whether fine-tuning was applied, and whether the parser was run with custom settings vs. default settings. In the table, the baseline configuration involves no use of a BERT model, no fine-tuning and an in-order as opposed to a top-down transition scheme.

The best overall F1 score was 90.38% in the case when we used IceBERT Base, fine-tuning and a top-down transition scheme (as opposed to the default in-order transition scheme). The parser achieved a slightly higher recall rate in the setup that was the same except that IceBERT large was used. This is a substantial improvement of the previous experiment in Arnardóttir and Ingason (2020). It is clear that the addition of an Icelandic BERT model makes a big difference for improving the results. The results mirror Snæbjarnarson et al. (2022) in that the best results are found with IceBERT base rather than IceBERT large, a somewhat surprising outcome. Snæbjarnarson et al. (2022) hypothesize that "this would change if the model is trained to convergence or better hyperparameter tuning".

Parser	IceBERT	Fine-tune	Top-down	F1 score	Precision	Recall
Berkeley Neural Parser				84.71%	85.07%	84.43%
(Arnardóttir & Ingason, 2020)						
Stanza						
Baseline	–	–	–	84.40%	84.92%	83.87%
IceBERT, Top-down	+Basic	–	+	88.66%	89.21%	88.11%
IceBERT, FT	+Basic	+	–	90.09%	90.32%	89.86%
IceBERT, FT, Top-down	+Basic	+	+	90.38%	90.54%	90.22%
IceBERT-Large, Top-down	+Large	–	+	88.80%	89.15%	88.46%
IceBERT-Large, FT	+Large	+	–	90.23%	90.40%	90.05%
IceBERT-Large, FT, Top-down	+Large	+	+	90.29%	90.39%	90.17%

Table 1: Results of the parsing experiment

5 Conclusion

In this paper we have described a new parsing setup for Icelandic that uses a CLARIN-available treebank for training and achieves better performance than the earlier system of Arnardóttir and Ingason (2020). The best F1 score in our experiments was 90.38% in the case when we used the IceBERT Base model, fine-tuning and a top-down transition scheme (as opposed to an in-order transition scheme). This is similar to the reported parsing accuracy in Snæbjarnarson et al. (2022), 90.02%, and slightly better, but the results are not directly comparable because different corpora were used for the experiments.

References

Arnardóttir, Þ., Hafsteinsson, H., Jasonarson, A., Ingason, A., & Steingrímsson, S. (2023, May). Evaluating a Universal Dependencies conversion pipeline for Icelandic. In T. Alumäe & M. Fishel (Eds.), *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 698–704). University of Tartu Library. https://aclanthology.org/2023.nodalida-1.69

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E. F., Bjarnadóttir, K., Ingason, A. K., Jónsdóttir, H., & Steingrímsson, S. (2020, December). A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In M.-C. de Marneffe, M. de Lhoneux, J. Nivre, & S. Schuster (Eds.), *Proceedings of the fourth workshop on universal dependencies (udw 2020)* (pp. 16–25). Association for Computational Linguistics. https://aclanthology.org/2020.udw-1.3

Arnardóttir, Þ., & Ingason, A. K. (2020). A neural parsing pipeline for icelandic using the berkeley neural parser. *Proceedings of CLARIN Annual Conference*, 48–51.

Ingason, A. K., Helgadóttir, S., Loftsson, H., & Rögnvaldsson, E. (2008). A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings*, 205–216.

Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., & Wallenberg, J. C. (2014, May). Rapid Deployment of Phrase Structure Parsing for Related Languages: A Case Study of Insular Scandinavian. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 91–95). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/855_Paper.pdf

Jökulsdóttir, T., Ingason, A., & Sigurðsson, E. (2019). A parsing pipeline for Icelandic based on the IcePaHC corpus. *Proceedings of CLARIN Annual Conference*, 138–141.

Jónsdóttir, H., & Ingason, A. K. (2020). Creating a parallel icelandic dependency treebank from raw text to universal dependencies. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2924–2931.

Kroch, A. S., Santorini, B., & Delfs, L. (2004). *Penn-Helsinki Parsed Corpus of Early Modern English. CD-ROM. First Edition.* [Size: 1.8 million words.].

Kroch, A. S., & Taylor, A. (2000). *Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition.* [Size: 1.3 million words.].

Loftsson, H. (2008). Tagging icelandic text: A linguistic rule-based approach. *Nordic Journal of Linguistics*, *31*(1), 47–72.

Loftsson, H., Bjarnadóttir, K., Helgadóttir, S., Whelpton, M., & Ingason, A. K. (2009). Icelandic language resources and technology: Status and prospects. *Nordic Perspectives on the CLARIN Infrastructure of Common Language Resources*, 27–33.

Loftsson, H., & Rögnvaldsson, E. (2007). Iceparser: An incremental finite-state parser for icelandic. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NoDaLiDa 2007)*, 128–135.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, *19*(2), 313–330.

Nikulásdóttir, A. B., Arnardóttir, Þ., Barkarson, S., Guðnason, J., Gunnarsson, Þ. D., Ingason, A. K., Jónsson, H. P., Loftsson, H., Óladóttir, H., Rögnvaldsson, E., et al. (2022). Help yourself from the buffet: National language technology infrastructure initiative on clarin-is. *CLARIN Annual Conference*, 109–125.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Rehm, G., & Way, A. (Eds.). (2023). *European language equality - A strategic agenda for digital language equality*. Springer. https://doi.org/10.1007/978-3-031-28819-7

Rögnvaldsson, E. (2010). Icelandic language technology: An overview. *Language, Languages and New Technologies: ICT in the Service of Languages. Contributions to the Annual Conference*, 187–195.

Rögnvaldsson, E., Ingason, A. K., & Sigurðsson, E. F. (2011). Coping with variation in the icelandic parsed historical corpus (icepahc). *Language Variation Infrastructure*, *3*, 97–112.

Rögnvaldsson, E., Ingason, A. K., Sigurðsson, E. F., & Wallenberg, J. (2012). The icelandic parsed historical corpus (icepahc). *LREC*, 1977–1984.

Sigurðsson, E. F., Ingason, A. K., Rögnvaldsson, E., & Wallenberg, J. C. (2012). Faroese parsed historical corpus (FarPaHC) 0.1 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/92

Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfsdóttir, S. L., Jónsson, H., Thorsteinsson, V., & Einarsson, H. (2022, June). A warm start and a clean crawled corpus - a recipe for good language models. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4356–4366). European Language Resources Association. https://aclanthology.org/2022.lrec-1.464

Wallenberg, J., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC)* [Version 0.9].

Þorsteinsson, V., Óladóttir, H., & Loftsson, H. (2019). A wide-coverage context-free grammar for icelandic and an accompanying parsing system. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 1397–1404.

˙<?xpacket begin="ïż£" id="W5M0MpCehiHzreSzNTczkc9d"?>˙<x:xmpmeta xmlns:x="adobe:ns:meta/">˙ <rdf:RDF x