CLARIN Annual Conference Proceedings

2020

Edited by

Costanza Navarretta, Maria Eskevich

05 – 07 October 2020
Virtual Edition

Please cite as:
CLARIN Annual Conference Proceedings, 2020. ISSN 2773-2177 (online).
Eds. C. Navarretta and M. Eskevich.
Virtual Edition, 2020.

Programme Committee

Chair:

- Costanza Navarretta, University of Copenhagen, Denmark

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Monica Monachini, Institute of Computational Linguistics “A. Zampolli" (IT)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eirikur Rögnvaldsson, University of Iceland (IS)
- 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)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eirikur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičenonienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Daan Broeder, NL
- Francesca Frontini, FR
- Maria Gavriilidou, GR
- Riccardo Del Gratta, IT
- Neeme Kahusk, EE
- Aleksei Kelli, EE
- Fahad Khan, IT
- Daniël de Kok, DE
- Penny Labropoulou, GR
- Christophe Parisse, FR
- Niccolò Pretto, IT
- Thorsten Trippel, DE
- Valeria Quochi, IT

CLARIN 2020 submissions, review process and acceptance

- Call for abstracts: 9 January 2020, 24 February 2020
- Submission deadline: 28 April 2020
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 15-16 June 2020
- Notifications to authors: 22 June 2020
- 36 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2020/clarin-annual-conference-2020-virtual-form.

Table of Contents

Repositories and Workflows

Data Curation, Archives and Libraries

Metadata and Legal Aspects

Extending the CLARIN Resource and Tool Families

Jakob Lenardič[1]
[1]Dept. of Translation, Faculty of Arts
University of Ljubljana, Slovenia
jakob.lenardic@ff.uni-lj.si

Darja Fišer[1,2]
[2]Dept. of Knowledge Technologies
Jožef Stefan Institute, Slovenia
darja.fiser@ff.uni-lj.si

Abstract

This paper presents the current state of the CLARIN Resource and Tool families initiative, the aim of which is to provide user-friendly overviews of the available language resources and tools in the CLARIN infrastructure for researchers from digital humanities, social sciences and human language technologies. The initiative now consists of a total of 11 corpus families, 5 families of lexical resource resources, and 4 tool families, which together amount to 950 manually curated tools and resources as of 17 August 2020. We present the initiative from the perspective of missing metadata as well as problems related to the general accessibility of the tools and resources and their findability in the Virtual Language Observatory (VLO).

1 Introduction

This paper presents the current state of the CLARIN Resource and Tool families initiative,[1] the aim of which is to provide user-friendly overviews of the available language resources and tools in the CLARIN infrastructure for researchers from digital humanities, social sciences and human language technologies. When the initiative first made its debut at LREC in 2018, it consisted of early versions of only 4 corpus families – parliamentary, computer-mediated communication (CMC), parallel, and newspaper corpora (Fišer et al., 2018). Since then, the initiative has been expanded substantially and now consists of a total of 11 corpus families, 5 families of lexical resources, and 4 tool families, which together amount to a manually curated list of 950 tools and resources as of 17 August 2020. Because one of our main aims in the initiative is to facilitate better curation of existing CLARIN tools and resources, this paper presents an overview of the identified issues with the metadata well as problems related to the findability and the accessibility of the resources and tools in the Virtual Language Observatory (VLO) (Van Uytvanck et al., 2012).

The paper is structured as follows. Section 2 presents a summary of the resource and tool families, focusing on their findability in the VLO, their accessibility, and missing metadata. Section 3 presents some of the current approaches to curating the families. Section 4 is the conclusion.

2 The CLARIN Resource and Tool Families

Tables 1 and 2 show that there are currently 558 individual corpora across the 11 surveyed corpus families, 283 resources across the 5 surveyed lexical resource families, and 109 tools across the 4 surveyed tool families. In the tables, we also provide summaries of the findability of resources and tools in the VLO, the degree of their accessibility, as well as missing metadata on size, annotation, and licence (for the resources) and input/output format and licence (for the tools).

2.1 Findability in the VLO and accessibility

The vast majority of the corpora (441 or 79% out of 558) as well as the lexical resources (268 or 95% out of 268) have VLO entries. The parallel and especially the L2-Learner corpus families are outliers

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

Family	**Items**	**VLO**		**Accessible**		**w/o size**		**w/o anno.**		**w/o licence**	
Parliamentary corpora	24	20	83%	12	96%	0	0%	3	13%	3	13%
CMC corpora	13	12	92%	13	100%	0	0%	2	15%	2	15%
Parallel corpora	86	53	62%	76	88%	11	13%	27	31%	6	7%
Newspaper corpora	33	22	67%	32	97%	2	6%	14	42%	3	9%
L2-learner corpora	74	34	46%	65	88%	8	11%	20	27%	4	5%
Historical corpora	73	66	90%	70	96%	12	16%	18	25%	6	8%
Spoken corpora	89	88	99%	80	90%	11	12%	25	28%	5	6%
Man. anno. corpora	73	68	93%	73	100%	1	1%	n/a	n/a	2	3%
Literary corpora	42	39	93%	39	92%	7	17%	23	55%	8	19%
Academic corpora	22	19	86%	21	96%	0	0%	12	55%	0	0%
Reference corpora	29	20	69%	28	97%	0	0%	3	10%	3	10%
Σ	558	441	79%	509	91%	52	9%	147	26%	42	8%
Lexica	75	72	96%	68	91%	2	3%	n/a	n/a	1	1%
Dictionaries	94	83	88%	93	99%	3	3%	n/a	n/a	19	20%
Conceptual resources	29	29	100%	27	93%	0	0%	n/a	n/a	1	3%
Glossaries	32	32	100%	32	100%	3	9%	n/a	n/a	1	3%
Wordlists	53	52	98%	53	100%	6	11%	n/a	n/a	4	8%
Σ	283	268	95%	273	96%	14	5%	n/a	n/a	26	9%

Table 1: The CLARIN Resource families, their findability in the VLO and their accessibility and missing metadata

with a significantly lower rate of findability in the VLO (62% and 46% respectively) because they contain a large number of resources that are provided either through the CLARIN:EL Central Inventory,[2] SWE-CLARIN's Språkbanken Text resource list,[3] or the SLABank L2 acquisition/learning collection in TalkBank,[4] none of which is currently harvested by the VLO.[5] We furthermore note the lower degree of findability of the tools in the VLO (80 or 73% out of 109) (Table 2), which is especially low for the normalizer family (4 or 29% of 14). This is likely due to the fact that the VLO search functionalities are currently tailored to language resources (Odijk, 2019).

Most of the corpora, lexical resources, and tools are readily accessible to the end user, generally for download through the CLARIN B-certified repositories or through CLARIN-related concordancers such as Korp (Borin et al., 2012) and Corpuscle (Meurer, 2012) in the case of corpora or web applications that are sometimes directly accessible through the CLARN repositories in the case of the tools, as with the *Morphodita morphological analyzer* (Straka and Straková, 2014),[6] while in certain cases, the repositories provide links to external pages where the tool or resource can be accessed.

Importantly, for the few inaccessible resources and tools, it is often the case that this is the result of seemingly neglected repository entries with underspecified documentation and outdated or irrelevant links to external landing pages. An example of this is the *One-million Corpus of Croatian Literary Language*,[7] which according to Tadić (2009, 220) is one of the precursors of the Croatian National Corpus. Its repository entry in LINDAT, which includes a limited description of the corpus with few metadata, offers only a hyperlink to a defunct project page,[8] so the corpus cannot be accessed.[9]

[2] `https://inventory.clarin.gr/`

[3] `https://spraakbanken.gu.se/resurser`

[4] `https://slabank.talkbank.org/`

[5] Note that TalkBank and CLARIN:EL, as well as the VLO software engineers, have been informed about the fact the VLO has problems harvesting their metadata, and are currently working on a solution to this problem. Given that the missing VLO entries are generally a global problem rather than an arbitrary quirk of individual resources, there will be a significantly higher rate of VLO inclusion in the Resource Families as soon as the CLARIN:EL and TalkBank repositories are correctly harvested.

[6] `http://lindat.mff.cuni.cz/services/morphodita/`

[7] `http://hdl.handle.net/11372/LRT-234`

[8] `http://hnk.ffzg.hr`

[9] Note that this corpus is a top-ranked search result in the VLO for the simple query "literary corpus"; cf. `https://vlo.`

2.2 Missing metadata

In this section we evaluate the resource and tool families from the perspective of their metadata description. Specifically, we note how many resources lack information on size, linguistic annotation (for the corpus families only), and licence, and how many tools lack information on input and/or output type as well as licence.

For the resources in Table 1, information on size and licence is generally readily provided across all the families, where size is not included only for 52 (9%) corpora and 14 (5%) lexical resources while licence is missing for 42 (8%) corpora and 26 (9%) lexical resources. It is noteworthy that an exception in terms of a high degree of missing licences is the dictionaries family. This family comprises a high number of Latvian dictionaries that are not yet included in the Latvian CLARIN repository, which in turn highlights the importance that resources be added to certified repositories conforming to e.g. FAIR principles (de Jong et al., 2018), as this would promote the inclusion of such metadata. By contrast, information on annotation in the case of the corpus families fares the worst, as it is missing for 26% of all the corpora. Including this information is crucial since it is otherwise difficult to distinguish corpora that are annotated but do not describe the annotation layers from corpora that are collections of unannotated plain text.

Family	**Items**	**VLO**		**Accessible**		**w/o format**		**w/o licence**	
Normalizers	14	4	29%	10	71%	5	36%	7	50%
NE recognizers	24	19	79%	23	96%	n/a	n/a	5	21%
Taggers and lemmatizers	66	54	82%	59	89%	41	62%	22	33%
Sentiment analysers	5	3	60%	4	80%	2	40%	1	20%
Σ	109	80	73%	96	88%	48	44%	35	32%

Table 2: The CLARIN Tool families, their findability in the VLO, their accessibility, and missing metadata

Finally, we note that the two types of observed metadata for the tools – i.e., input/output format and licence in Table 2 – are the least readily included types of information. For instance, licence is missing for 32% of all the tools, and is especially seldom included for the text normalizer family (missing for half of the surveyed normalizers). Interestingly, the text normalizers also constitute the family that has the least number of VLO entries, which further emphasises the importance of releasing tools through B-certified repositories that facilitate a thorough documentation of the metadata. Relatedly, input and output formats remain undocumented for 44% of all the tools, which is again likely tied to Odijk (2019)'s observation that the majority of the currently used CMDI-metadata profiles are tailored to data rather than software.

3 Curating the Families

Tables 1 and 2 list a total of 364 issues pertaining to missing metadata on annotation, size, licence, or input/output formats. One of the main goals of the CLARIN Resource and Tool Families initiative is to promote better documentation of CLARIN's resources and tools. As a first step towards this we have compiled the issues on a GitHub page[10] and classified them according to the type of missing metadata (unclear annotation, missing licence, etc.). With the help of national CLARIN representatives, we have thus far managed to solve a total of 84 (19%) such metadata issues. We have also assigned labels specifying the CLARIN consortium that is responsible or most closely associated with the issue to all the GitHub entries, thereby further incentivizing individual CLARIN consortia to help with the curation process. Certain issues have also been successfully resolved by directly referring them back to their original data contributors.

In the future, we will also evaluate the resource and tool families from a more qualitative perspective, taking into account not only the provision of metadata, but also in which way metadata are reported for the observed categories. For instance, it is often the case that resource size is reported in broadly different

clarin.eu/?0&q=literary+corpus.

[10] https://github.com/clarin-eric/resource-families-issues

ways even for resources within the same resource family. For example, where some corpora in the same family report size in terms of sentences, others report it in terms of tokens, words, hours or file size,[11] which hinders the cross-comparability of the resources. In addition to different measurement units for size, certain resources specify their annotation layers very imprecisely, using vague descriptors such as "multitagged",[12] without further qualifying them, which is uninformative and not user-friendly. In the case of licence information, some tools and resources list unhelpful values such as an "other" licence, which is problematic from the perspective of the mapping to the CLARIN licence categorisation (e.g., CLARIN PUB, CLARIN ACA)[13] displayed in the VLO. To help counter this in the future, we will also work on developing and promoting detailed guidelines for the depositing of new resources and tools, thereby maximising their usability by the research community.

4 Conclusion

We have presented the current state of the CLARIN Resource and Tool Families initiative, describing it mainly from the perspective of missing metadata and problematic aspects of accessibility and findability in the VLO. A concentrated and increased effort in the curation of the families is of crucial importance because existing issues hinder the findability and reuse of any resources and tools in the CLARIN infrastructure but especially those featured in the CLARIN Resource and Tool Families, which has proven to be a highly visible initiative appreciated by a broad spectrum CLARIN users and therefore warrants continued and careful upkeep.

Acknowledgements

We would like to thank the CLARIN National Coordinators as well as the members of the User Involvement Committee, whose help and input has always been crucial for expanding and maintaining the CLARIN Resource and Tool Families. We would also like to thank the anonymous reviewers for their helpful comments.

References

[Andersen2014] Gisle Andersen. 2014. *The Norwegian Newspaper Corpus*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11372/LRT-370.

[Borin et al.2012] Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 474–478.

[Carrasco Benitez2013] Manuel Tomas Carrasco Benitez. 2013. *Official Journal of the European Union_oj4fd2*. clarin:el. http://hdl.grnet.gr/11500/ATHENA-0000-0000-23E1-A.

[de Jong et al.2018] Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259–3264.

[Fišer et al.2018] Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1320–1325.

[Meurer2012] Paul Meurer. 2012. Corpuscle – a new corpus management platform for annotated corpora. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 29–50. Benjamins, Amsterdam. http://hdl.handle.net/10.1075/scl.49.02meu.

[11] E.g., the *Official Journal of the European Union* corpus (Carrasco Benitez, 2013)

[12] E.g., *The Norwegian Newspaper Corpus* (Andersen, 2014)

[13] https://www.clarin.eu/content/licenses-and-clarin-categories

[Odijk2019] Jan Odijk. 2019. Discovering software resources in clarin. In *Selected Papers of the CLARIN Conference 2018*, volume 159, pages 121–132. Linköping University Electronic Press, Linköpings universitet.

[Straka and Straková2014] Milan Straka and Jana Straková. 2014. *MorphoDiTa: Morphological Dictionary and Tagger*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. `http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0`.

[Tadić2009] Marko Tadić. 2009. New version of the croatian national corpus. In *After half a century of Slavonic natural language processing*, pages 199–205. Tribun EU.

[Van Uytvanck et al.2012] Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 1029–1034. European Language Resources Association (ELRA).

An internationally FAIR Mediated Digital Discourse Corpus: towards scientific and pedagogical reuse

Rachel Panckhurst
Dipralang EA 739
Université Paul-Valéry Montpellier 3
Montpellier, France
rachel.panckhurst@univ-montp3.fr

Francesca Frontini
Praxiling UMR 5267 CNRS
Université Paul-Valéry Montpellier 3
Montpellier, France
francesca.frontini@univ-montp3.fr

Abstract

In this paper, the authors present a French Mediated Digital Discourse corpus, (*88milSMS* http://88milsms.huma-num.fr https://hdl.handle.net/11403/comere/cmr-88milsms). Efforts were undertaken over the years to ensure its publication according to the best practices and standards of the community, thus guaranteeing compliance with FAIR principles and CLARIN recommendations with pertinent scientific and pedagogical reuse.

1 Introduction

The adoption of Open Data and Open Science principles is producing important effects in SSH disciplines and has been enhanced by widespread awareness of the internationally ratified FAIR principles, aiming at ensuring that research data should be Findable, Accessible, Interoperable and Reusable[1]. In Linguistics and Natural Language Processing (NLP) the importance of curating Language Resources (LRs) for replicability and reuse has been particularly recognized, with relevant initiatives dating back several decades. Even before the formalisation of the FAIR principles as such, various initiatives have promoted good data management practices in the domain of Language Resources, starting with the FLaReNet[2] action and culminating with the creation of the CLARIN infrastructure. With its network of consortia and centres, CLARIN is making it easier for researchers to adhere to the requirements of the FAIR principles (de Jong et al., 2018), something which is increasingly required by evaluation and funding agencies. In France, over and above its role as CLARIN observer, various national centres are now active under the leadership of the national Huma-Num infrastructure, offering services and promoting the sharing of textual data (among other types) which meet the FAIR principles and the CLARIN best practices.

European Computer-Mediated Communication (CMC) and Mediated Digital Discourse (MDD) corpora initiatives are becoming more visible: Belgian *sms4science*, *Vos Pouces*, (Fairon et al., 2006; Cougnon, 2015; Cougnon and Fairon, 2014; Cougnon et al., 2017); Dutch SoNaR, (Oostdijk et al., 2008); French CoMeRe, (Chanier et al., 2014); German DeRik, (Beißwenger et al., 2013); Swiss *What's up Switzerland?*, (Ueberwasser and Stark, 2017; Frey et al., 2016). These data types are often difficult to process, standardize, analyze, owing to their complex nature, including 'noisy' content (Frey et al., 2019; Poudat et al., 2020).

The objective of this paper is to present *88milSMS*, a French CMC/MDD corpus; in particular, we shall highlight the efforts undertaken over the years to ensure its publication according to the best practices and standards of the community, which in turn guarantee compliance with FAIR principles and CLARIN recommendations. Scientific and pedagogical reuse is emphasized.

[1] http://datafairport.org/fair-principles-living-document-menu

[2] http://www.flarenet.eu/

2 Project & Corpus

The *sud4science* project[3] was part of a vast international initiative, entitled *sms4science*[4], which aimed at building a worldwide database and analysing authentic text messages in different languages — mainly French, but also Creole, German (written in Switzerland and Germany), Italian, Romansh (Dürscheid and Stark, 2011), and English (Drouin and Guilbault, 2016). Many scientific projects analyse authentic data, but ensuing corpora are not always made available for the scientific community and the general public, sometimes owing to legal requirements and commercial isses. However, there is a crucial need for researchers from a wide range of disciplines to have easy access to authentic data, in order to conduct analyses pertaining to their particular research fields. From the onset of the *sud4science* project, the possibility of easy access and reuse of authentic data was of utmost importance to the scientific team.

In 2011, over 88,000 authentic French text messages were collected during a 13-week period from the general public in Montpellier, France (Panckhurst et al., 2016b) and SMS 'donors' were also invited to fill out a sociolinguistic questionnaire (Panckhurst and Moïse, 2014). An anonymization phase was conducted (Patel et al., 2013), owing to legal requirements for data-protection of private data (Ghliss and André, 2017). This involved anonymizing names, telephone numbers, places, brand names, addresses, codes, URLs[5]. In 2014, the finalised largest digital resource of 88,000 'raw' anonymized French text messages, the specific *88milSMS* corpus, two samples (1,000 transcoded SMS, 100 annotated SMS), and the sociolinguistic questionnaire data were made available for all[6] to download. The researchers chose the Huma-Num web service[7], then in 2016, they made a TEI/XML version of *88milSMS* available under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence on the 'Ortolang' platform, which provided the corpus with a citable persistent identifier[8]. Contributions to DARIAH and ELRA were also made in 2015, and *88milSMS* therefore has an ISLRN. These initiatives preceded the FAIR principles but are in strict alignment with them.

3 Towards being FAIR

Findability refers to initiatives aimed at ensuring long-term preservation of LRs by depositing in a specialized data centre (Ortolang repository, France), documented by a rich and standardized set of metadata, which in turn can be harvested by international meta-catalogues (CLARIN Virtual Language Observatory), allowing international visibility. Thanks to the deposit on the Ortolang repository, the *88milSMS* corpus is **Findable** from the VLO, where it will appear by performing free text searches such as "cmc corpus" or "SMS corpus" and filtering by language. In addition to the visibility within the VLO, the corpus is indexed on Google, searchable on the ELRA catalogue, as well as on Isidore[9].

Accessibility is gained by adopting a clear set of licences and promoting open access within copyright limits and data protection regulations; single-sign-on technologies allow researchers to gain access to resources based on their institutional identifier. *88milSMS* is fully **accessible**, despite its sensitive content, thanks to the thorough anonymization and verification work (Ghliss and André, 2017) carried out with the help of the University's legal advisors; a short mandatory form needs to be completed to download from the bilingual (French/English) Huma-Num web interface, with a user free-of-charge licence. However, no authentication or form completion are needed to download the corpus from Ortolang, where it is available via the Creative Commons CC BY 4.0 licence.

Interoperability and **re-usability** for LRs are particularly important, and crucially enabled by the use of standard annotation formats and common best practices, allowing researchers to exploit data from

[3] `http://sud4science.org`; (Panckhurst et al., 2016b).

[4] `http://www.sms4science.org`; (Fairon et al., 2006; Cougnon and Fairon, 2014; Cougnon, 2015).

[5] By default, first names, surnames and any data which enable identifying information are anonymized. It is of course frustrating for linguists and other scientists to feel that anonymization causes loss of information, which will not be able to be retieved at a later stage, but it is a stringent legal requirement.

[6] Both the scientific community and the general public.

[7] `http://88milsms.huma-num.fr` (Panckhurst et al., 2014), .

[8] `https://hdl.handle.net/11403/comere/cmr-88milsms` (Panckhurst et al., 2016a).

[9] Isidore is a French search engine for documents and resources in SSH `https://isidore.science/document/http://hdl.handle.net/11403/COMERE/V3.3/CMR-88MILSMS`

different projects. At the time of the data collection, similar initiatives took place at the international level (*cf.* § 1) thus making the overall philosophy of the corpus attuned to that of these other datasets. Indeed, other authentic data collections projects followed on in more recent years (Ueberwasser & Stark 2017; Cougnon *et al*. 2017). From the point of view of encoding, initial formats of the corpus were .ods spreadsheets and *ad hoc* .xml. The use of utf-8 was crucial at the time (2011), in particular to ensure the preservation of a subset of SMS containing the first instances of emoji (Panckhurst and Frontini, 2020). The work to make *88milSMS* fully **interoperable** was carried out later, with the inclusion within the CoMeRe initiative, where the project adopted a common TEI format. Thanks to the aforementioned efforts, *88milSMS* has been **reused** beyond the initial scope of the project, boasting more than 1,000 downloads from 51 countries, and with a broad spectrum of multidisciplinary applications (to name a few: language sciences, computational linguistics and text-mining processing initiatives, geographical place name identification, psychology case studies).

4 Towards scientific and pedagogical reuse

Three years after providing *88milSMS* for public download and dissemination, a survey on scientific usage of the corpus was conducted[10]. Results have shown a strong disciplinary tendency towards language sciences and computing including NLP, text mining and corpus linguistics research, mainly from higher education establishments. In terms of dissemination, 50% of the research cited was successfully circulated in Master's theses, PhDs, habilitations, books, articles, proceedings, etc. (Panckhurst et al., 2020). In 2019, an update survey was conducted in order to find out if colleagues had cited/used *88milSMS*. Responses were unfortunately minimal – and the authors need to improve the way in which corpus reuse information is obtained – but they do indicate that the corpus is being used in language sciences, as is to be expected, but also in other disciplines:

1. Language Sciences and NLP:
    - university courses for 2nd-year students; identifying and improving spelling mistakes (Poitiers University); discourse genres (Lorraine University);
    - recent PhDs: French as a foreign language and how to include SMS-writing in didactic situations; linguistic analysis of French SMS-writing; SMS communication: NLP and information extraction;
    - qualitative comparative analysis between differing corpora, related to morphosyntactic French question-form usage and interactional aspects comparing SMS and oral language.
2. Geography: identification of place names and interpretation of variations (Master's 2 internship subject, 2019, IGN-Paris & Paris-Est Marne-la-Vallée University).
3. Psychology: digital communication and teenagers (relational, emotional romantic aspects, 12-16 year-olds, Master's 1 thesis 2019, Toulouse Jean-Jaurès University).

5 Conclusion and future work

In this paper, the authors indicated how the French Mediated Digital Discourse *88milSMS* corpus complies with the four FAIR principles (findability, accessibility, interopeability and reusability), also taking into account CLARIN recommendations.

Even though the corpus is fairly widely consulted, downloaded and used across the scientific community and beyond, it remains difficult to have sufficient access to other researchers' scientific and pedagogical reuse, despite implementation of an optional scientific newsletter and surveys.

The authors' own pedagogical usage at Université Paul-Valéry Montpellier 3 (under-graduate and post-graduate levels) for Language Science students includes studying discourse analysis and NLP techniques with contemporary instant messaging authentical data such as the *88milSMS* corpus, which has recently been incorporated on the widely consulted Sketch Engine (`http://www.sketchengine.eu/`)

[10]Researchers and the general public who had signed up to an optional scientific newsletter were contacted.

platform, thus allowing online analysis via a user-friendly interface without mandatory downloading. The advantage of this sort of integration is to provide ever-increasing interdisciplinary scientific and pedagogical reuse possibilities.

In this sense, the collaboration on CMC corpora which has already started within CLARIN is crucial[11] for the harmonization of formats across international projects, for the identification of common technical solutions for browsing interfaces, and finally for the implementation of a Federated Content Search.

The next step is long-term archiving of *88milSMS* and other FAIR corpora at the National Computing Center for Higher Education (CINES)[12], providing insight on digital textuality usage for future generations.

References

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. 2013. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 28(4):531–537. Publisher: Oxford Academic, https://academic.oup.com/dsh/article/28/4/531/1077484.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J., and Seddah, D. 2014. The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *JLCL - Journal for Language Technology and Computational Linguistics*, 29(2):1–30. https://halshs.archives-ouvertes.fr/halshs-00953507.

Cougnon, L.-A. and Fairon, C., editors. 2014. *SMS Communication*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Cougnon, L.-A., Maskens, L., Roekhaut, S., and Fairon, C. 2017. Social media, spontaneous writing and dictation. Spelling variation. *Journal of French Language Studies*, 27(3):309–327. https://www.cambridge.org/core/journals/journal-of-french-language-studies/article/div-classtitlesocial-media-spontaneous-writing-and-dictation-spelling-variationdiv/9574CD6BF604BD8F866A270E1EC909A3.

Cougnon, L.-A. 2015. *Langage et sms: Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. 2018. CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1515.

Drouin, P. and Guilbault, C. 2016. De 'Viens watcher la partie avec moi' à 'Come regarder the game with me'. In *Abstracts, PLIN 2016*, Louvain-la-Neuve, Belgium.

Dürscheid, C. and Stark, E. 2011. sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. In Crispin, T. and Mroczek, K., editors, *Digital Discourse. Language in the New Media*. Oxford University Press. ISBN: 9780199795437, https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199795437.001.0001/acprof-9780199795437-chapter-14.

Fairon, C., Klein, J. R., and Paumier, S. 2006. *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*. Presses universitaires de Louvain. Manuel.CD-Rom., Louvain-la-Neuve.

Frey, J.-C., Glaznieks, A., and Stemle, E. W. 2016. The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In Corazza, A., Montemagni, S., and Semeraro, G., editors, *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016. 5-6 December 2016 Napoli*, pages 157–161, Torino. Academia University Press. https://bia.unibz.it/handle/10863/8949.

Frey, J.-C., König, A., and Stemle, E. W. 2019. How FAIR are CMC corpora? In Longhi, J. and Marinica, C., editors, *Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (cmccorpora19), Cergy-Pontoise University, France, 9-10 September 2019*, pages 26–31. https://bia.unibz.it/handle/10863/11294.

[11] See for instance the past thematic event `https://www.clarin.eu/event/2017/clarin-plus-workshop-creation-and-use-social-media-resources` as well as the CMC 'Resource Family' entry `https://www.clarin.eu/resource-families/cmc-corpora`

[12] `https://www.cines.fr/en/`

Ghliss, Y. and André, F. 2017. Après la collecte, l'anonymisation : enjeux éthiques et juridiques dans la constitution du corpus 88milSMS. In Ciara R. Wigham, G. L., editor, *Corpus de Communication Médiée par les Réseaux*, pages 71–84. L'Harmattan, Paris. https://hal.archives-ouvertes.fr/hal-01722169.

Oostdijk, N., Reynaert, M., Monachesi, P., Noord, G. V., Ordelman, R., Schuurman, I., and Vandeghinste, V. 2008. From D-Coi to SoNaR: a reference corpus for Dutch. In *Proceedings of LREC 2008*, Marrakech, Morocco. ELRA. http://www.lrec-conf.org/proceedings/lrec2008/pdf/365_paper.pdf.

Panckhurst, R. and Frontini, F. 2020. Evolving interactional practices of emoji in text messages. In Thurlow, C., Dürscheid, C., and Diémoz, F., editors, *Visualizing Digital Discourse. Interactional, Institutional and Ideological Perspectives*, pages 81–103. De Gruyter Mouton.

Panckhurst, R. and Moïse, C. 2014. French text messages. From SMS data collection to preliminary analysis. In Cougnon, L.-A. and Fairon, C., editors, *SMS Communication. A Linguistic Approach*, pages 141–168. John Benjamins. https://hal.archives-ouvertes.fr/hal-01485595.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. 2014. 88milSMS. A corpus of authentic text messages in French, produit par l'Université Paul-Valéry Montpellier III et le CNRS, en collaboration avec l'Université catholique de Louvain, financé grâce au soutien de la MSH-M et du Ministère de la Culture (Délégation générale à la langue française et aux langues de France) et avec la participation de Praxiling, Lirmm, Lidilem, Tetis, Viseo. ISLRN : 024-713-187-947-8, https://hal.archives-ouvertes.fr/hal-01485560.

Panckhurst, R., Détrie, C., Lopez, C., Moïse, C., Roche, M., and Verine, B. 2016a. 88milSMS. A corpus of authentic text messages in French. In Chanier, T., editor, *Banque de corpus CoMeRe*. Nancy, France. Ortolang, https://hdl.handle.net/11403/comere/cmr-88milsms.

Panckhurst, R., Roche, M., Lopez, C., Verine, B., Détrie, C., and Moïse, C. 2016b. De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire. *Histoire Epistémologie Langage*, 38(2):63–82. https://hal.archives-ouvertes.fr/hal-01485577.

Panckhurst, R., Lopez, C., and Roche, M. 2020. A French text-message corpus: 88milSMS. Synthesis and usage. *Corpus [online]*, (20). http://journals.openedition.org/corpus/4852.

Patel, N., Accorsi, P., Inkpen, D., Lopez, C., and Roche, M. 2013. Approaches of anonymisation of an SMS corpus. In *Proceedings of CICLING 2013, LNCS*, pages 77–88, March 24-30, 2013, University of the Aegean, Samos, Greece. Springer-Verlag. https://hal-lirmm.ccsd.cnrs.fr/lirmm-00816285.

Poudat, C., Wigham, C. R., and Liégeois, L. 2020. *Corpus complexes. Traitements, standardisation et analyse des corpus de communication médiée par les réseaux*. Corpus (20).

Ueberwasser, S. and Stark, E. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5). https://bop.unibe.ch/linguistik-online/article/view/3849.

The First Dictionaries in Esperanto. Towards the Creation of a Parallel Corpus

Denis Eckert
Géographie-Cités / Centre Marc Bloch
CNRS, Paris / Berlin
eckert@parisgeo.cnrs.fr

Francesca Frontini
Laboratoire Praxiling - CNRS
Université Paul-Valéry Montpellier 3
francesca.frontini@univ-montp3.fr

Abstract

Between 1887 - date of creation of the Esperanto language by L. Zamenhof in Warsaw - and 1890, 37 documents in at least 16 different languages were prepared by Zamenhof himself or by early adopters of Esperanto, in order to present the new "international language" to the broadest possible public. This unique multilingual corpus is scattered across many national libraries. We have systematically collected digital versions of this set of documents and begun to analyze them in parallel. Many of them (17) contain the same basic dictionary, elaborated by people who mostly had a limited, or absolutely no knowledge of philology. These 17 dictionaries encompass about 920 entries in Esperanto, each time translated in a given target language. We are progressively digitizing the whole corpus of these small dictionaries (so far, 12 versions have been digitized and encoded) and aim at making it accessible to scholars of various disciplines in TEI format in an Open Access repository. The need for international and interdisciplinary co-operation is obvious, due to the great variety of the languages used, in order to decipher and correctly encode linguistic issues that are likely to arise (non-standardized Hebrew, pre-independence Latvian or Lithuanian, old-style Russian spelling, etc.).

1 Introduction

By mid-1887 in Warsaw - then part of the Russian Empire, an anonymous Dr Esperanto published a brochure presenting a new language called by him *Lingvo internacia* - the international language. The author, Ludwik Lazar Zamenhof (1859-1917), was an ophthalmologist, born in Białystok to a Jewish family of intellectuals (O'Keeffe, 2019; Korzhenkov, 2009).

Though Esperanto, as the new language was soon named, never became the universal language for international communication it was meant to be, it is still in use worldwide. Esperanto is the only constructed language that has not disappeared several years after its birth (as *Volapük*, *Ido*, *Langue bleue* and many other similar projects did). One of the resilience factors of this language is probably that it is easy to learn (a classical argument of its activists). But organizational factors most certainly also played a big role (Forster, 1982). Esperantists set up an international organization at a very early stage; and they were able to hold their first World Congress in 1905 (Boulogne-sur-Mer).

Prior to all these organizational efforts, the first explantory factor of the early international diffusion of Esperanto is the quick translation/adaptation of the initial brochure in many languages. Within five months (from June until November 1887), the *First Book* or *Unua Libro*, as the brochure was nicknamed in the milieu of the early adopters, had been published in four main European languages (first in Russian, then in Polish, French, and German, see Figure 1). The sole author/translator of the four versions is Ludwik Zamenhof. The following year (1888), four new brochures were released, this time not prepared by the creator. The *First Book* was translated into English, Hebrew, and Yiddish by two Warsaw acquaintances of Zamenhof; and, by the end of the year, the first systematic textbook (for German speakers) was published in Nuremberg by a local adopter. During the next years, many adaptations of the initial brochure were released, targeting people from different nationalities.

The *First Book* as such (i.e. the early version of the document) is both a manifesto for an international language and a short description of the basic principles of Esperanto, with a small dictionary of around 920 entries (words or basic roots) in Esperanto and their translation in the target language. With time, the contents of the succeeding versions evolved, abandoning *peu à peu* the style of the initial manifesto to become more and more similar to standard textbooks for foreign languages.

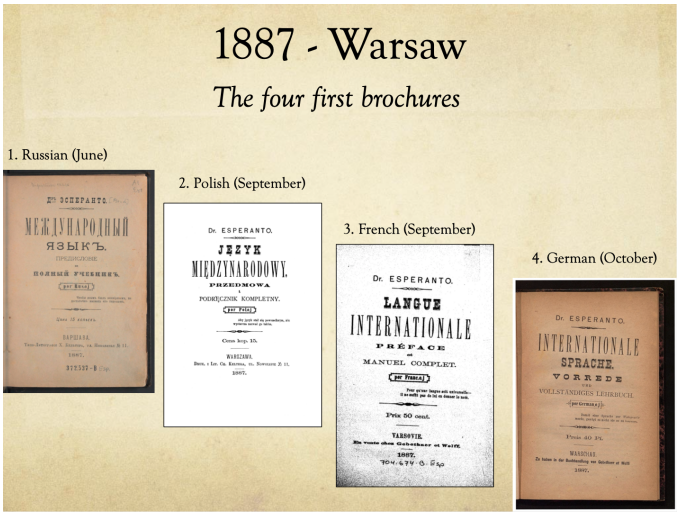


Figure 1: The first four editions of the *Unua Libro*, in Russian, Polish, French and German.

To sum up, by the end of 1890, 45 documents approved by L. Zamenhof had been published: the *First Book* itself, textbooks, dictionaries, essays; and texts in Esperanto (including poetry and literary translations). They are all included in the list of approved publications maintained throughout decades by the head of the Esperantist movement[1]. Eight of these documents are exclusively in Esperanto, thus addressing the audience of early adopters, while the 37 others, which aimed at presenting Esperanto to the international public, are written in 16 different languages.

- Languages used for presenting Esperanto (1887 – 1890) in the first booklets and textbooks:

  Russian, Polish, French, German, English, Hebrew, Yiddish, Swedish, Spanish, Latvian, Romanian, Danish, Bulgarian, Italian, Czech, Lithuanian.

2 The project and the corpus

In this corpus, dictionaries are abundant (more than 20). 17 of these dictionaries deserve, in our view, particular attention.

We have gathered electronic facsimiles of all publications containing a dictionary. While many of these documents are stored (and digitized in image format) at the *Department of Planned Languages* of the Österreichische Nationalbibliothek in Vienna, a substantial number of rare items appeared to be scattered accross Europe, the USA and Israel. They were, when necessary, digitized upon our request (2019 – 2020).

By examining these documents, we could identify a specific sub-corpus: the first Esperanto-natural languages dictionaries. Indeed, many brochures and textbooks encompass the same small dictionary of Esperanto words (around 920 entries). This limited dictionary (henceforth called *vortaro*, plural *vortaroj*) is composed of basic roots that can be used in verbs, adjectives, adverbs, etc. It is most frequently a uni-directional dictionary (Esperanto-target language), according to the template of the four initial brochures

[1]The list is called *Nomaro pri l'verkoj pri la lingvo internacia Esperanto* (list of works related to the Esperanto international language). Each publication has a serial number, below Esperanto Serial Number or ESN.

of 1887[2]. Surprisingly, some languages were provided with more than one *vortaro* (up to 3). This can be explained by the fact that, in some cases, different authors prepared brochures separately, targeting the same linguistic group. The total number of comparable dictionaries amounts to 18 (in 15 languages). For analytical reasons, we exclude the second German *vortaro*[3], thus reducing the corpus to 17 items. The list of the 17 editions containing the dictionaries can be found in Table 1; a more complete version of the same table is available in the online documentation[4].

ESN	Target language	Year	Author / Translator		Library	Place
6	Russian	1887	Zamenhof	Ludwik L.	ÖNB	Vienna
7	Polish	1887	Zamenhof	Ludwik L.	ÖNB	Vienna
8	French	1887	Zamenhof	Ludwik L.	ÖNB	Vienna
9	German	1887	Zamenhof	Ludwik L.	ÖNB	Vienna
10.1	English	1888	Steinhaus	Julius	BL	London
10.2	English	1889	Geoghegan	Richard	ÖNB	Vienna
13	Hebrew	1888	Najmanovich	Naftali	NLI	Jerusalem
22	English	1889	Phillips	Henry	LoC	Washington
23	Yiddish	1888	Najmanovich	Naftali	ÖNB	Vienna
26	Spanish	1889	de Wahl	Edgar	ÖNB	Vienna
27	Swedish	1889	Henriclundquist	Gustav	ÖNB	Vienna
29	Latvian	1889	Libeks	Rudolf	ÖNB	Vienna
32	Romanian	1889	Frollo	Marietta	BML	Lyon
34	Bulgarian	1890	Bogdanov	Mikhail	BNL	Sofia
35	Italian	1890	Marignoni	Daniele	ÖNB	Vienna
39	Czech	1890	Lorenc	Ferenc	ÖNB	Vienna
47	Lithuanian	1890	Dombrovski[5]	K. A.	ÖNB	Vienna

Table 1: List of the 17 editions. ESN stands for Esperanto Serial Number

The characteristics of the corpus:

- All these dictionaries, with small variations, share the same list of Esperanto roots (from 914 to 921 entries).
- They were, with one single exception[6], prepared by non-specialists, as a voluntary (and unpaid) contribution to the development of the Esperantist movement.
- Dictionaries were mostly prepared in a non-centralized and non-coordinated way by people who lived in different cities (Warsaw, Philadelphia, London, Saint-Petersburg, Crema, Lund, Bucarest, Riga, Kaunas, Sofia, etc.).
- A high degree of similarity exists between dictionaries, with few differences in the lists of entries.

Because of these peculiar aspects, these dictionaries constitute an original and interesting parallel corpus, which has so far neither been put together nor analyzed.

3 The publication

The project currently concentrates on the publication of the *vortaroj*, but a publication of the complete editions is envisaged. The working documents, as well as a table summing up the progresses, are available online[7]. Currently all existent *vortaroj* have been collected in image format and visually inspected. Twelve of them (in 10 different languages) have been fully digitized and are available in .csv format.

[2]Such rudimentary *vortaroj* ceased to be published after 1893. All later dictionaries encompass a much bigger number of entries, and are systematically bi-directional.

[3]This *vortaro*, part of a textbook published by L. Einstein in Nürnberg (1888), is a simple replication of the first German *vortaro* published the year before by Zamenhof in Warsaw.

[4]See working paper at `https://halshs.archives-ouvertes.fr/halshs-02555912v1`

[5]The Polish orthograph is used in the title page; the Lithuanian orthograph "Dabrauskas" is used to sign the preface.

[6]R. Geoghegan, author of one of the three English versions, had studied philology in Oxford, but never graduated.

[7]`https://hal.archives-ouvertes.fr/ESPERANTO-HISTORIO`

The next step in the project is the conversion into TEI, using the lexicographic module. In creating the data model we have adhered as closely as possible to the guidelines of TEI-lex0 (Bański et al., 2017). The TEI version will faithfully describe the content of each *vortaro*, with precise metadata information about the original edition; at the same time it should allow and facilitate the exploration of the collection, by means of an interlinguistic alignment of each lexical entry as well as by enriching the corpus with explicit linguistic information, such as part of speech. Currently, a proposal of TEI modelling has been made for two *vortaroj* (French and German).

4 Links to CLARIN

We believe that the corpus can be of interest for scholars studying the evolution of both Esperanto and of the various target languages, in particular as concerns previous stages of orthography (e.g. old style Russian, Hebrew, Baltic languages) and different scripts (Latvian). In view of this, a collaboration with CLARIN researchers from various countries could help in:

- finding and digitizing further editions,
- establishing partnerships and collaborations for the encoding of specific languages and editions,
- consolidating the choice of data model and annotations,
- providing visibility to the corpus (which will be made available via a CLARIN center) and encouraging linguistic analysis,
- offering guidance as to copyright issues.

5 Conclusion

The encoding of the *vortaroj* is the first step of a more comprehensive project. The early years of Esperanto constitute a remarkable moment of creation of a universalist contributing project in the age of "First Globalization" (Berger, 2003). It should therefore be systematically documented. Locating and gathering copies of all these documents (some of them being extremely rare) is, in our view, not enough. In order to make this corpus accessible to a great variety of scholars (historians, linguists, digital humanists, cultural studies scholars, etc.), our perspective is to achieve a full-text digitization of all the early works (1887 – 1891) of the Esperanto movement, and make them accessible online in a standardized format[8]. Besides, our project is not just for researchers. We also believe that civil society can be interested in these documents, especially the global community of Esperantists, who will be able to have a documented look at the early history of the language of which they are activists. For that purpose, the constitution of an international multidisciplinary network, with adequate funding, appears to be the only suitable answer to such a challenge.

References

Bański, P., Bowers, J., and Erjavec, T. 2017. TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In *eLex2017*.

Berger, S. 2003. *Notre première globalisation: leçons d'un échec oublié*. Seuil, Paris.

Forster, P. G. 1982. *The Esperanto movement*. Number 32 in Contributions to the sociology of language. Mouton, The Hague ; New York.

Korzhenkov, A. 2009. *Zamenhof: The Life, Works and Ideas of the Author of Esperanto*. Mondial - Universal Esperanto Association.

O'Keeffe, B. 2019. An International Language for an Empire of Humanity: L. L. Zamenhof and the Imperial Russian Origins of Esperanto. *East European Jewish Affairs*, 49(1):1–19.

[8]Thus avoiding the difficult question of the copyrights issues related to digitized items belonging to national libraries.

Digital Neuropsychological Tests and Biomarkers: *Resources for NLP and AI Exploration in the Neuropsychological Domain*

Dimitrios Kokkinakis
Department of Swedish
University of Gothenburg,
Sweden
dimitrios.kokinakis@gu.se

Kristina Lundholm Fors
Department of Health & Rehabilitation
University of Gothenburg,
Sweden
kristina.lundholmfors@gu.se

Abstract

Non-invasive, time and cost-effective, easy-to-measure techniques for the early diagnosis or monitoring the progression of brain and mental disorders are at the forefront of recent research in this field. Natural Language Processing and Artificial Intelligence can play an important role in supporting and enhancing data driven approaches to improve the accuracy of prediction and classification. However, large datasets of e.g. recorded speech in the domain of cognitive health are limited. To improve the performance of existing models we need to train them on larger datasets, which could raise the accuracy of clinical diagnosis, and contribute to the detection of early signs at scale. In this paper, we outline our ongoing work to collect such data from a large population in order to support and conduct future research for modelling speech and language features in a cross-disciplinary manner. The final goal is to explore and combine linguistic with multimodal biomarkers from the same population and compare hybrid models that could increase the predictive accuracy of the algorithms that operate on them.

1 Introduction

According to WHO (2019) over 50 million people worldwide have been diagnosed with dementia, and there are nearly 10 million new cases annually. Dementia, a gradual decline in cognitive function due to neurodegeneration, is a growing concern as the global population ages. No cure is available, but research suggests that early intervention may lead to better outcomes for both individuals and their caregivers, achieving greater impact of the available treatment options (Posner et al., 2017). To enable early intervention, we must improve our ability to identify people with subtle signs of cognitive decline, i.e. Mild Cognitive Impairment (MCI); cf. Petersen et al. (2014). MCI is characterized by clinically observable deficits in at least one cognitive domain, but where the deficits do not interfere with a person's ability to complete the activities of daily living. Approximately half the persons who develop MCI will later develop dementia. The final goal of this research is to examine how early cognitive decline affects language in relation to biomarkers. To achieve this, we aim to apply methods from Natural Language Processing (NLP) and Artificial Intelligence (AI; incl. Machine Learning, ML & Deep Learning, DL). The data to use is a multimodal and multidimensional dataset gathered in the H70-1944 Birth Cohort Study in Gothenburg. A subset of this material, the digital neuropsychological assessment, will be used for the linguistic exploration. The language-based results will then be combined with results from other parallel biomarker studies of the same cohort. We hypothesize that such a combination could improve the accuracy of the differential diagnosis of neurodegenerative conditions. Our vision is to better understand and manage neurodegenerative impairment early, investigate correlations between features and modalities, improve classification and prediction models and make differential diagnosis more accurate.

2 Background from a Language Technology Perspective

Cognitive health is an important determinant of functional ability in older adults (Laske et al., 2015; Graham et al., 2020). When trying to detect the early stages of cognitive decline, the complexity of

language makes finding meaningful patterns a great challenge, while conventional neuropsychological tests may not be sensitive to some of the earliest changes. Advances in NLP & AI technologies have contributed to great progress for the identification of early signs of cognitive deterioration. Recent studies employ algorithms that model multimodal features in AI architectures. E.g., Beltrachini et al. (2015) and others have shown that the detection of early cognitive decline can be improved by combining features from MRI with cognitive test scores in ML. Similarly, Fraser et al. (2019) showed that fusing data from multiple modalities improved the discriminative performance of the classifier and outperformed a classifier trained on a single modality. Recent research has also indicated that automated language analysis for the extraction and modelling of linguistic features for automated diagnostic models for e.g. Alzheimer's Disease (AD) improves steadily (Mueller et al., 2017; Filiou et al., 2019). Progress in employing language as a rich source of information regarding the individual's cognitive status has the potential to advance our understanding of the early stages and the progress of dementia and its impact on language (Meilán et al., 2014; Fraser et al., 2016).

To improve the overall performance of classification algorithms on linguistic features, we need to train models on larger datasets. Small datasets do not generalize well (Masrani et al., 2017) while large language-related datasets in this domain are scarce, cf. Li et al. (2019) for a review. The most widely used datasets for linguistic feature extraction are: (i) the *TalkBank*, which includes the *DementiaBank* (Becker et al., 1994; Orimaye et al., 2017); (ii) the *Framingham Heart Study* (Satizabal et al., 2016); and the (iii) the *Wisconsin Registry for Alzheimer's Prevention* (Johnson et al., 2018).

3 Participants, Data and Methods

NLP involves transferring language from an unstructured format into a structured one to enable analyses in learning frameworks. For clinicians, language plays a central role in diagnosis (Ahmed et al., 2013) and, recently, statistical and NLP-supported AI have been applied in this domain that could potentially lead to technological breakthroughs for early detection of cognitive decline. This section provides more information on the data we collect and the methodological approach we aim to apply.

3.1 Participants

The participants in this study belong to the H70-1944 cohort, which is part of the Gothenburg birth cohort studies (H70; Rydberg Sterner et al., 2019). The H70 studies, carried out at the 'Centre for Ageing and Health' (AgeCap) at the U. of Gothenburg, are some of the largest longest-running epidemiological studies on aging and dementia in the world. The H70-1944 cohort (70-year-olds born 1944, n=1202) was examined in 2014-16. All participants were assessed at the memory clinic in Gothenburg. The cohort is re-examined in 2019-20 and undergoes digital neuropsychological assessment for the first time. Since collection of the data is half-way at the time of the writing (n=410), we cannot provide exact figures on the categorization of the participants to controls and those who belong to major MCI subtypes such as amnestic or non-amnesic according to established diagnostic procedures (DSM-5). Expert assessment/ground truth of the cohort will be provided during 2021. Note that due to the COVID-19 pandemic, the study has been temporarily suspended and it is estimated to restart during fall 2020.

3.2 Digital Assessment and Empirical Data

All H70-1944-participants complete a battery of computerized, neuropsychological assessments. The assessment is digitally recorded using an iPad platform ('Δelta iPad app', ki elements, Germany); all language-based tests are also recorded using high quality audio (44.1 kHz sampling freq.; 16 bit resolution). Digital assessment implies that not only the spoken signal is registered but also the participant's movements of a digital pen across the iPad's screen (used in some of the tests). The cognitive battery includes tests covering the cognitive domains: memory, language, executive functions, visuospatial skills and speed/attention. For example, the *Boston Naming Test* (BNT), a confrontational naming test, measures word retrieval; the *Verbal Fluency* tests (VF) measures lexical access and executive control ability (semantic and phonemic VF). The *Rey Auditory Verbal Learning Test* (RAVLT) evaluates e.g. short-term auditory-verbal memory, rate of learning, confabulation of confusion in memory processes, and differences between learning and retrieval. The *Stroop* test evaluates cognitive interference resolution as well as assesses the influence of selective attention on information processing speed. The *Trail Making Test A&B* provides information about e.g. speed of processing, and executive functioning.

3.3 Biomarkers

Lumbar puncture (CSF) is up to now performed in 26% of the H70-1944 cohort (*n*=322) and baseline data for CSF levels of NFL, Neurogranin, Aß42, T-tau and P-tau and brain MRI data is available. Genetic data is available for approximately 1160 individuals from the H70-1944 cohort. By using the weightings for a large number of genetic variants, based on publicly available genome wide association studies (GWAS) data, genotyped individuals can have Polygenic Risk Scores (PRS) calculated that measure the cumulative small effects producing genetic vulnerability of the disorder. Structural and functional MRI are performed in about half of all participants. Volumetric and connectivity measures are derived from scans performed on a 3T Philips scanner. Tau PET using the 18F-RO948 ligand is performed in a subsample of individuals with previously acquired MRI and CSF data (*n*=100). Semi-quantitative standardized uptake values will be derived for all anatomical brain regions.

3.4 Linguistic and Digital Markers

The use of digital methods increases the reliability of cognitive assessments and the possibility to acquire, test and evaluate a considerable amount of new markers. This will allow for greater breadth and depth when exploring speech and language data than we have shown in our previous research, where we utilized a much smaller cohort: see Fraser et al. (2018; 2019); Themistocleous et al. (2020). We will apply both NLP and signal processing methods to extract clinically meaningful features from the individuals' speech samples, including *prosodic features* (e.g. fundamental frequency), *articulatory features* (e.g. Mel-Frequency Cepstral Coefficients), *voice quality* (e.g. Harmonic-to-Noise Ratio), and *paralinguistic features* (e.g. fillers, pause length and speech-rate). Features from the transcribed speech will include *lexical features* (e.g. frequency of lexical items, word repetitions), *morphosyntactic features* (e.g. complexity of lexical items such as compounds vs non-compounds), and *semantic features* (e.g. lexical and semantic similarity). Moreover, the iPad usage and digital pen (cf. 3.2) enables extraction of new types of features (e.g. duration of pen corrections). The linguistic and digital data will be the source of experimentation and from which we will perform correlational feature analysis with biomarkers (cf. 3.3). Since manual transcription of speech is costly and time-consuming, we applied automatic speech recognition using Google's Automatic Speech Recognition (ASR) service for the completed recordings. This allows us to also extract features from text, i.e. the transcriptions, and use the research infrastructure developed within the National Language Bank of Sweden ("Nationella språkbanken") for processing the data. For feature extraction from speech, we plan to use the *openSMILE* package (Eyben et al., 2013).

3.5 Data Processing

AI uses computational algorithms to analyze complex structured and unstructured data (e.g. speech and text). AI tools can use high-dimensional data to determine e.g. potential predictors of normal versus pathological changes in cognitive functioning. The performance depends on the model selected, available data, and the input features used to predict an outcome. We will train different models and assess their performance separately and combined, and we will investigate the jointly most predictive feature sets and the types of errors the AI models make. Iteratively, we will improve models' performance through parameter fine-tuning using cross-validation. The magnitude of the data is important for a comprehensive and reliable interpretation of the factor analysis results. Neural networks will be implemented using standard packages for neural network design such as *KERAS* (Chollet et al., 2015).

4 Future Directions and Challenges: Linguistic and Digital Biomarkers

There is a growing interest and need to extract and use digitally assessed linguistic biomarkers for the assessment of disease severity, complications and prognosis of brain and mental disorders. Compared to classical pen-and-paper tests, the use of linguistically based and digital biomarkers have many advantages providing non-invasive, unobtrusive measurements of cognitive health. Recent evidence suggests that combining results from different technologies may improve the accuracy for both the prediction of cognitive decline and the detection of e.g. dementia. LT and AI tools use large volumes of multifeature data to determine potential predictors of normal versus pathological changes in cognitive functioning. Digitally captured features are also less prone to human bias. For instance, acoustic features extracted from the recordings, e.g. salient features such as the frequency and duration of pauses and voice quality features such as fundamental frequency, can unlock previously unobtainable sources of

behavioural, social, and physiological variation with minimal effort required from the participants and the medical experts. Previous research has shown that learning linguistic biomarkers from the utterances of individuals could provide important complementary information to help the clinical diagnosis of various early onset neurodegenerative diseases. However, there is a need to train the predictive and diagnostic models on larger datasets. In the majority of previous studies, the sample sizes were too small to draw safe conclusions about the general population; the H70-1944 cohort is a population from which the experiments conducted pave the way for promising, future explorations and advances into the understanding of early signs of cognitive impairment with greater accuracy.

Acknowledgements

This work has received support from *the Swedish Foundation for Humanities & Social Sciences* (grant: NHS 14-1761:1) and from *the Centre for Ageing and Health* at the U. of Gothenburg (AgeCap).

References

Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*. 136 (Pt 12):3727-3737.

James T. Becker, François Boiler, Oscar L. Lopez, et al. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arc of Neur*. 51:6, 585-594.

Leandro Beltrachini, Matteo De Marco, Zeike A. Taylor, et al. 2015 Integration of Cognitive Tests and Resting State fMRI for the Individual Identification of MCI. *Curr Alzheimer Res.* 12(6):592-603.

François Chollet, et al. 2015. Keras. github.com/kerasteam/.

Florian Eyben, Felix Weninger, Florian Gross and Björn W Schuller. 2013. *Recent developments in openSMILE, the Munich open-source multimedia feature extractor*. 21st ACM Conf on Multimedia. 835-38. Spain.

Renée-Pier Filiou, Nathalie Bier, Antoine Slegers, et al. 2019. Connected speech assessment in the early detection of Alzheimer's disease and mild cognitive impairment: a scoping review. *Aphasiology*. 34:6, 723-755.

Kathleen C Fraser, Jed A. Meltzer and Frank Rudzicz. 2016. Linguistic features identify AD in narrative speech. *J of Alz Dis*. 49:2.

Kathleen C Fraser, Kristina Lundholm Fors and Dimitrios Kokkinakis. 2018. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *J Comput Speech Lang*. 53: 121-139.

Kathleen C Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman and Dimitrios Kokkinakis. 2019. Predicting MCI Status from Multimodal Lang. Data using Cascaded Classifiers. *Front Aging Neurosc*. 11:205.

Sarah A. Graham, Ellen E. Lee, Dilip V. Jeste, et al. 2020. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psych Res*. Vol. 284.

Sterling C. Johnson, Rebecca L. Koscika, Erin M. Jonaitis, et al. 2018. The Wisconsin Registry for AD Prevention. *Alz&Dem*. 10:130-42.

Christoph Laske, Hamid R Sohrabi, Shaun M Frost, et al. 2015. Innovative diagnostic tools for early detection of AD. *Alz&Dem.* 11(5).

Yiling Li, Yi Lin, Hongwei Ding & Chunbo Li. 2019. Speech databases for mental disorders. *BMJ Gen Psyc.* 32:3.

Vaden Masrani, Gabriel Murray, Thalia Shoshana Field, and Giuseppe Carenini. 2017. *Domain adaptation for detecting MCI.* AI Conf. Springer, 248–259.

Juan J.G. Meilán, Francisco Martínez-Sánchez, Juan Carro, et al. 2014. Speech in Alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia & Geriatric Cogn Dis*. 37(5-6): 327–334.

Kimberly Mueller, Rebecca L. Koscik, Bruce Hermann, et al. 2017. Declines in connected language are associated with very early MCI: Results from the Wisconsin registry for Alz prevention. *Front in aging neurosc*. 9:437.

Sylvester O. Orimaye, Jojo S-M. Wong, Karen J. Golden, et al. 2017. Predicting probable AD using linguistic deficits and biomarkers. *BMC Bioinf.* 18:34.

Ronald Petersen, Barbara Caracciolo, C. Brayne, et al. 2014. MCI: a concept in evolution. *J Int Med.* 275:214-28.

Charalambos Themistocleous, Marie Eckerström & Dimitrios Kokkinakis. 2020. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PLOS ONE* 15(7):e0236009.

Therese Rydberg Sterner, Felicia Ahlner, Kaj Blennow, et al. 2019. The Gothenburg H70 Birth cohort study 2014-16: design, methods and study population. *Eur J Epidemiol*. 34(2):191-209.

Holly Posner, Rosie Curiel, Chris Edgar, et al. 2017. Outcomes assessment in clinical trials of AD and its precursors: readying for short-term and long-term clinical trial needs. *Innov. Clin. Neurosci.* 14:22.

Claudia L. Satizabal, Alexa S. Beiser, Vincent Chouraki, et al. 2016. Incidence of dementia over three decades in the Framingham Heart Study. *N. Engl. J. Med.* 374, 523–532.

WHO. 2019. Dementia: www.who.int/news-room/fact-sheets/detail/dementia. Visited 2020/02/19.

CORLI: The French Knowledge-Centre

Eva Soroli
University of Lille, CNRS, UMR 8163, MESHS, France
efstathia.soroli@univ-lille.fr

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

Flora Badin
LLL-CNRS, Orléans, France
flora.badin@univ-orleans.fr

Antonio Balvet
University of Lille, CNRS, UMR 8163, MESHS, France
antonio.balvet@univ-lille.fr

Elisabeth Delais-Roussarie
University of Nantes, CNRS UMR 6310-LLING, France
elisabeth.delais-roussarie@univ-nantes.fr

Carole Etienne
ICAR-CNRS, Lyon, France
carole.etienne@ens-lyon.fr

Lydia-Mai Ho-Dac
CLLE, University of Toulouse CNRS, UT2J, France
hodac@univ-tlse2.fr

Loïc Liégeois
University of Paris, France
loic.liegeois@univ-paris-diderot.fr

Christophe Parisse
INSERM, CNRS-Paris, Nanterre University, France
cparisse@parisnanterre.fr

Abstract

As a first step towards increasing reproducibility of language data and promoting scientific synergies and transparency, CORLI (Corpus, Language and Interaction), a consortium involving members from more than 20 research labs and 15 Universities, part of the French large infrastructure Huma-Num, contributes to the European research infrastructure of CLARIN through the establishment of a knowledge sharing centre: the French Clarin CORpus Language and Interactions K-Centre (CORLI K-Centre). The purpose of the CORLI K-Centre is to provide expertise in corpus linguistics and French language, and support academic communities through actions towards FAIR and Open data. We describe the development of the CORLI K-Centre, its scope, targeted audiences, as well as its intuitive and interactive online platform which centralizes and offers both proactive and reactive services about: available language resources, databases and depositories, training opportunities, and best research practices (i.e., on legal/ethical issues, data collection, metadata standardization, anonymization, annotation and analysis guidelines, corpus exploration methods, format conversions and interoperability).

1 Introduction

More and more researchers underline the need to give data greater value, make them digital and interoperable as well as enhance their propensity for reuse. As a step towards increasing reproducibility of the data and promoting scientific collaboration and transparency, a group of researchers (Wilkinson et al. 2016) postulated the so-called FAIR principles (making data findable, accessible, interoperable and reusable). Such guiding principles are relevant for any scientific discipline but also increasingly relevant for linguistics, especially in the fields of corpus linguistics, natural language processing and computational linguistics typically characterized by great variability, isolated data collection practices, heterogeneous formats, etc. (Ciamiano et al. 2020).

In order to alleviate such issues related to incompatibility and promote interoperability, we propose to contribute to the European research infrastructure of CLARIN through the establishment of a French Clarin Knowledge Centre, the CORLI K-Centre, in the domain of corpus linguistics and French language based on the French CORLI (Corpus, Language, and Interaction) consortium (Parisse et al. 2018).

The aim of this paper is to share our experience and our network in developing a new K-Centre and to benefit from knowledge and recommendations from existing European K-Centres.

1.1 France, CLARIN, Huma-Num and the consortia

In the last decade, the French government and the CNRS (French National Research Institute) have funded several projects and initiatives to provide methods and tools to researchers working on digital humanities. Data centres dedicated to storage and dissemination of language data have been created, such as ORTOLANG and COCOON (both browsable centres in the CLARIN VLO), aiming at providing access to databases within sustainable digital platforms.

Huma-Num is currently the main player in the field. Supported by the CNRS and the French Ministry of Research, Huma-Num is a large research infrastructure with international reach devoted to Social Sciences and Humanities that promotes through technical support and funds research networks called *consortia*. CORLI is one of these consortia. Founded in 2016, CORLI provides services similar to those offered by CLARIN K-centres, and is dedicated to consensus-based recommendations and digital solutions in corpus linguistics.

1.2 The French CORLI consortium

CORLI is an open consortium of more than fifteen Universities, twenty laboratories and hundreds of researchers around France specialized in corpus linguistics and covering a very wide set of theoretical approaches (previously exposed by Parisse et al. 2018). CORLI's primary role is to be a reference centre for corpora and language resources, and a network that creates and distributes a wide array of language data, tools, methods and digital solutions. More specifically, we promote methodological approaches for corpora, from their collection to their storage, and we have become instrumental in promoting good practices, digital tools and standards by encouraging researchers to make their methods, outcomes and data open, accessible and reusable. As an incentive, we: (a) provide financial help to standardize and complete already existing corpora (oral, written, multimodal) through a yearly call for proposals; (b) closely monitor the development of new data; (c) support training sessions in digital annotation and analysis tools; and (d) we are attentive to the needs of the community, especially in domains such as metadata standardization, anonymization, corpus exploration guidelines, format conversions, interoperability as well as legal and ethical matters.

2 CORLI as a K-Centre

Although most of our goals are close to what other CLARIN K-Centres do, creating a new K-Centre calls for some clarification of our actions and specificities.

2.1 K-Centres and the CORLI K-Centre's expertise

As mentioned above, the *CORLI K-Centre* has two main areas of expertise: (a) corpus linguistics; and (b) the French language. With respect to corpus linguistics, and unlike other K-Centres specialized on specific topics such as computer-mediated communication, multimodality, sign language or learner corpora (i.e. the *K-Centre for Atypical Communication Expertise*), our centre's scope is closer to what more general centres do, such as the *CKLD: CLARIN Knowledge-Centre for linguistic diversity and language documentation* and *the Norwegian Centre for Research Data*, in that we offer expertise on data collection and data processing, as well as assistance for corpus building, annotation and data management.

With respect to the French language and its varieties, our Centre has similar goals to those of the *Czech CLARIN K-Centre for Corpus Linguistics*– structured around the Czech National Corpus and covering a wide range of actions (e.g., centralisation and mapping of language data and resources) – in that we offer centralized information about existing national corpora, annotation manuals and guidelines for the analysis of the French language and the languages of France to researchers working in this domain or interested in crosslinguistic comparisons.

The CORLI K-Centre is specialized, however not limited, to French language corpora, and thus can provide expertise on any type of language resources and language technology, as well as training opportunities for acquiring maximum autonomy in building and sharing language data – actions that dovetail with the objectives of other centres, such as the *CLARIN Knowledge Centre for South Slavic languages (CLASSLA)* and the *CLARIN K-Centre DANSK - DANish helpdeSK*.

2.2 Objectives and actions of the CORLI K-Centre

The main ambition of the CORLI K-Centre is to achieve a transformation of the research lifecycle in corpus linguistics and French language studies: offer expert advice from a panel of experienced

investigators about available databases and digital tools, provide resources to enhance the quality and reporting of linguistic and related research, support junior and early stage researchers in their training and development, and encourage FAIR data creation, edition and reuse.

In order to achieve these goals, the CORLI K-Centre functions as an intuitive and interactive on-line platform (already available from: https://corli.huma-num.fr/en/kcentre) which centralizes and provides cross-border access to knowledge through both proactive and reactive services.

With respect to proactive knowledge sharing, the CORLI K-Centre offers through a thematically organized website information about: (a) data sharing and access, (b) metadata standardization procedures, (c) format conversion and available software for language processing, (d) corpus exploration methods and tools, (e) guidelines and available manuals for corpus annotation, (f) legal issues related to corpus management and use and (g) training opportunities. The development of a FAQ (Frequently asked questions) page addressing common concerns in these topics (e.g. copyright, research ethics, research design, data collection, automatic analysis, corpus exploration methods) will further contribute to information access. The users of the CORLI K-Centre platform will have the possibility to access most knowledge through the website of the centre, and alternatively through the FAQ, where other landing pages will offer the possibility to redirect to related content (e.g., to ERIC, CLARIN, TalkBank, etc.) and thus continue the journey ideally without the need for outside assistance.

In cases of requests for further assistance, the CORLI K-Centre offers an additional reactive knowledge-sharing service established thanks to a pool of researchers and data specialists who can provide further information if needed. The way the users will interact with the webpage and the provided knowledge is of vital importance to the CORLI group. For this reason, a contact form has been integrated to the platform (easily accessible from a separate Contact-Us button) offering the possibility to the users who cannot find the answer to their questions to contact the centre directly.

3 Methodological challenges and solutions

Some of the greatest challenges in corpus linguistics are related to the great variability of practices as well as to the diversity of the corpora themselves (Cox 2011). With respect to the nature of the corpora, language data, by definition, present a huge variability: some researchers work with written data, others with spoken or gesture data; some focus on child language, others on adult use; some are interested in special populations and case studies (bilinguals/multilinguals, people with language disorders), others on natural language processing using very large corpora. With respect to the practices, investigators, according to their research questions and targeted populations, often opt for isolated data collection practices, incompatible annotation systems/formats and variable (in-house) management, storage and analysis methods that only rarely address ethical issues or allow for sharing and reuse. In addition, irrespective of types and formats and officially since 2018, researchers are invited to provide detailed information on the procedures for data collection, storage, protection, archiving and destruction of the collected data and thus conform to the General Data Protection Regulation (GDPR).[1]

As a first step towards increasing interoperability, transparency, protection and reproducibility of language data, CORLI has been managing, for several years now, six working groups that address these challenges. More specifically, the groups follow a committee approach, and produce consensus-based guidelines and recommendations in the following areas: Interoperability, query and annotation tools; Multimodality and new forms of communication; Multilingualism; Legal issues and Data protection; Best practices for corpus annotation; and Corpus assessments.[2]

The expertise gathered by the pool of specialists involved in the above groups has led to a great amount of outcomes, useful to linguists (all levels of academic expertise) but also to anyone working with corpora or interested in language use, databases, digital tools for data exploration and data management (engineers, data scientists, educators, etc.). Based on the outcomes of previous and current work, a series of knowledge sharing documents and tools are developed that can be used for serving the CLARIN community at large and thus meet the needs of a broader audience (e.g., Table 1 below).

[1] The GDPR is directly applicable in the EU Member States since 25 May 2018. The full text is available here

[2] For further information about the activities of our working groups see: https://corli.huma-num.fr/en/projectgroups

Resources	Purpose
Knowledge management diagrams	Data collection steps (video/audio data, constraints, to-do lists)
	The corpus lifecycle (including iterative processes, archiving and reuse)
	Flow diagram for corpus development and annotation methodology
Practice papers & Guidelines	Best practice recommendations for metadata, format conversions and sustainability
	Guidelines for anonymization and data protection
	Guidance to students, researchers, authors, editors and publishers about proper citation of language datasets, research projects, new annotations in archived corpora etc.
Technical manuals & useful documents	Templates of informed consents
	Manuals for transcription tasks with most commonly used software (Clan, Praat, Elan, Transcriber etc.) and minimal transcription recommendations
	Manuals: research methods and analysis (e.g., computer-mediated communication, sign language, multilingual corpora)
Bloopers	Typical errors in annotated corpora, in metadata etc.
	Most common anonymization issues
	Examples of typical speech errors in corpora (e.g., special populations, child data)

Table 1. Examples of useful knowledge sharing documents

The development of a K-Centre based on the CORLI consortium will bring acquired consensus-based expertise, services and digital solutions to a broader audience, will facilitate the creation of international synergies among actors interested in language corpora and French linguistics, and thus strengthen the participation of France (currently an observer) to the CLARIN infrastructure.

4 Conclusion

To conclude, the work of the groups on recommendations, good practices, tools and methods, as well as the activities and commitments of the network to help researchers facing new challenges (GDPR, open science, data management), make CORLI the appropriate organization for a CLARIN K-Centre. With expertise in corpus linguistics and French language, the CORLI K-Centre aims to become a major platform of knowledge sharing and communication among researchers and other actors interested in language and corpus linguistics (data scientists, engineers, educators, etc.). Built on the shared knowledge of the CORLI consortium, the new CORLI K-Centre will benefit from our past experience and current projects in tool development and practices, and contribute with actions that fit perfectly with the scope of other CLARIN K-Centres. The centre will provide through both proactive and reactive services useful information about available national and international databases and depositories, manuals and annotation techniques for French and other languages, corpus exploration methods, conversion platforms for interoperability, advice on legal issues, metadata standardization procedures, webinars and online training opportunities, and thus will contribute to international synergies and enhance knowledge and practice sharing.

References

Cimiano, Ph., Chiarcos, Ch., McCrae, J. & Gracia, J. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

Cox, C. 2011. Corpus linguistics and language documentation: challenges for collaboration. In J. Newman, J., Baayen, H. & Rice S. (eds.) *Corpus-based studies in language use, language learning and language documentation* (p. 239-264). Brill, Rodopi.

Parisse, C., Poudat, C., Wigham, C. R., Jacobson, M., & Liégeois, L. 2018. CORLI: A linguistic consortium for corpus, language, and interaction. In *Selected papers from the CLARIN Annual Conference 2017*, Budapest, 18–20 September 2017 (p. 15-24). Linköping University Electronic Press, Linköpings universitet.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data 3*: 160018.

The CLASSLA Knowledge Centre for South Slavic Languages

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

Petya Osenova
IICT-BAS
Sofia, Bulgaria
petya@bultreebank.org

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

Kiril Simov
IICT-BAS
Sofia, Bulgaria
kivs@bultreebank.org

Abstract

We describe the recently set-up CLARIN Knowledge centre CLASSLA focused on language resources and technologies for South Slavic languages. The Knowledge centre is currently run by the Slovene national consortium CLARIN.SI and the Bulgarian national consortium CLADA-BG. The two main aims of the Knowledge centre are coordination in development of language resources and technologies for the typologically related languages and joint training activities and helpdesk support for the underlying user base.

1 Introduction

Knowledge centres in the CLARIN infrastructure are organizational instances consisting of one or multiple CLARIN national consortia that excel in a specific topic and offer their expertise to other national consortia and the users of the infrastructure. Currently there are 23 CLARIN certified knowledge centres.[1]

The South Slavic language group consists of seven official languages: Bosnian, Bulgarian, Croatian, Macedonian, Montenegrin, Serbian and Slovene. This language group is traditionally divided into Western South Slavic languages (Bosnian, Croatian, Montenegrin, Serbian and Slovene, all of which, except Slovene, are considered to be mutually intelligible) and Eastern South Slavic languages (Bulgarian and Macedonian, being for the most part mutually intelligible).

There were three primary motivations for setting up the CLASSLA Knowledge centre on South Slavic languages:

1. The Slovene national consortium had a strong track record in developing language resources and technologies (LRTs) not only for Slovene, but also for Croatian and Serbian (the latter two being for the most part applicable to Bosnian and Montenegrin as well), while the Bulgarian national consortium has a strong record in developing LRTs for Bulgarian (which are partially applicable to Macedonian).

2. The availability of LRTs for South Slavic languages is rather low as documented in the META-NET white papers for four out of the seven countries in which South Slavic languages are official languages (Blagoeva et al., 2012; Tadić et al., 2012; Vitas et al., 2012; Krek, 2012) and organising the Knowledge centre should foster collaboration among LRT developers for this language group.

3. Only three out of the seven countries have their CLARIN national consortia set-up[2] and this Knowledge centre could be a good entry point for the remaining four languages into the CLARIN ERIC infrastructure.

[1] The full list of knowledge centres can be found at https://www.clarin.eu/content/knowledge-centres.

[2] At the moment the Knowledge centre started being organised, only Slovenia and Bulgaria had their national consortia set-up. Croatia was at that point a candidate and became in the meantime full member of CLARIN ERIC. We hope that CLARIN-HR will join the CLASSLA Knowledge centre as well.

The potential for a synergistic development of LRTs for South Slavic languages can be exemplified with the recent activities of collaborative development of training corpora for Slovene, Croatian and Serbian on the levels of morphosyntax, dependency syntax, social media language processing, named entity recognition and semantic role labeling, described in detail in (Ljubešić et al., 2018b).

The CLASSLA Knowledge centre has received a CLARIN ERIC certificate and official Knowledge centre status on 19 March 2019.

2 Overview of the CLASSLA Knowledge Centre Resources

In this section we give an overview of the various resources (people, data, technologies) the CLASSLA Knowledge centre currently offers.

2.1 Helpdesk

The CLASSLA helpdesk, the most prominent feature of the CLASSLA website[3], offers information on the existing LRTs for South Slavic languages, as well as assistance in developing LRTs. Some examples of the LRTs recently developed with the assistance of the Knowledge centre are:

- The first corpus of the Montenegrin language (a parallel English–Montenegrin corpus consisting of subtitles from the Montenegrin television) (Božović et al., 2018b; Božović et al., 2018a).
- A web corpus of the Montenegrin language (meWaC) consisting of 91 million tokens.
- Tokenizers for the Bulgarian and Macedonian language included in the `reldi-tokeniser` tool[4] which already supported Croatian, Serbian and Slovene.
- The `classla-stanfordnlp` tool[5][6] (fork of the `stanfordnlp` natural language processing pipeline) for annotating text on the levels of morphosyntax, lemmas, dependency syntax and named entities, the pipeline currently covering Slovene, Croatian, Serbian, and Bulgarian.

The near-future plans of the helpdesk include processing the Bulgarian Reference Corpus (Simov and Osenova, 2008) with the `classla-stanfordnlp` pipeline and including it in the CLARIN.SI concordancers (described in Section 2.3) and adding at least partial support for Macedonian to the `classla-stanfordnlp` pipeline.

2.2 Frequently Asked Questions

The entry point to the CLASSLA Knowledge centre for a new language is setting up an FAQ document on the respective language and its LRTs. Currently the Knowledge centre covers FAQs for Bulgarian, Croatian, Serbian, and Slovene.There are plans of writing an FAQ for Macedonian by the Knowledge centre partners from Northern Macedonia.

2.3 Concordancers

The Knowledge centre currently offers two CLARIN.SI concordancers, the NoSketch Engine (Rychlý, 2007) and KonText (Machálek, 2020). The concordancers have the same back-end and cover the same corpora and it is up to preferences or specific requirements of the researcher to choose one of them. The number of corpora for the languages of the Knowledge centre are currently the following (in decreasing order): Slovene (52), Croatian (7), Serbian (4), Montenegrin (2), Bosnian (1), Bulgarian (1), Macedonian (0). While the prevalence of Slovene corpora is not surprising, as the concordancers form part of the Slovenian CLARIN.SI infrastructure, the amount of corpora for the other languages is far from negligible and it will surely improve in the years to come.

[3] `https://www.clarin.si/info/k-centre/`
[4] `https://github.com/clarinsi/reldi-tokeniser`
[5] `https://github.com/clarinsi/classla-stanfordnlp`
[6] `https://pypi.org/project/classla/`

2.4 Data Repository

The Knowledge centre also uses the CLARIN.SI repository. The number of entries for the languages of the Knowledge centre are currently the following (in decreasing order): Slovene (133), Croatian (26), Serbian (21), Bulgarian (11), Bosnian (3), Macedonian (3) and Montenegrin (1). Similar as with concordancers, the prevalence of Slovene entries does not come as a surprise, but we expect the nuber of resources for the least represented languages to steadily improve.

Apart from Slovene, there are also some crucial ones for Croatian and Serbian, e.g., the largest freely available training datasets of standard language, hr500k (Ljubešić et al., 2018a) and SETimes.SR (Batanović et al., 2018)) and the Internet (non-standard) language (ReLDI-NormTagNER-hr (Ljubešić et al., 2019a) and ReLDI-NormTagNER-sr (Ljubešić et al., 2019b)) and the overall largest corpora (hrWaC (Ljubešić and Klubička, 2016a) and srWaC (Ljubešić and Klubička, 2016b)) and inflectional lexicons (hrLex (Ljubešić, 2019a) and srLex (Ljubešić, 2019b)). available for these two languages.

2.5 Web Services

The web services of the Knowledge centre currently hold the name ReLDIanno since their initial version was developed inside the ReLDI project[7]. The services have a web interface[8] and some corresponding documentation[9] both on the web interface as well as a Python library that can be used to invoke the web services programmatically. The services currently offer support for Croatian, Serbian and Slovene, processing running text on the levels of tokenization, sentence splitting, tagging, parsing and named entity recognition. The web application also offers batch processing, i.e., uploading zip files consisting of multiple text files, all being processed with the selected procedure.

While the back-end of the current web services are old, pre-neural technologies (Slovene morphosyntactic annotation obtains on these an F1 score of 94.21 while the recent neural state-of-the-art achieves F1 of 97.06[10]), version 2.0 of ReLDIanno is being developed which will have the neural `classla-stanfordnlp` pipeline as a back-end (described in 2.1), providing state-of-the-art results for the languages of the Knowledge centre, as well as enabling adding new languages (Macedonian) and additional varieties (standard language, non-standard language, Serbo-Croatian macro-language etc.).

2.6 Training Activities

The first CLASSLA workshop was to be held in Ljubljana in 2020, but had to be, due to the COVID-19 crisis, partially postponed and partially held online. We hope this to be the first of many CLASSLA workshops aiming to (1) disseminate the CLASSLA LRTs to interested parties, but also (2) identify partners interested in developing new or improving existing LRTs for South Slavic languages.

3 Conclusion

In this abstract we have presented the CLASSLA Knowledge centre on South Slavic languages and its already significant capacity given the short period this Knowledge centre exists.

Apart from documenting at least some of the resources of the target languages, the aim of this publication is also to motivate other researchers and national consortia in setting up similar knowledge centres aimed not only at sharing expertise, but also organising collaborative efforts in developing and improving LRTs for less-resourced languages and language groups.

Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments and suggestions and the numerous collaborators of the K-centre, especially those whose recent activities helped increase the LRT coverage of the K-centre: Vuk Batanović, Petar Božović, Maja Miličević Petrović, Aleksandar Petrovski, Biljana Stojanovska, Tanja Samardžić, Milica Vuković Stamatović and Katerina Zdravkova.

[7] `https://reldi.spur.uzh.ch/`
[8] `http://clarin.si/services/web/query`
[9] `http://www.clarin.si/info/k-centre/web-services-documentation/`
[10] `https://github.com/clarinsi/babushka-bench#morphosyntactic-tagging`

References

[Batanović et al.2018] Vuk Batanović, Nikola Ljubešić, Tanja Samardžić, and Tomaž Erjavec. 2018. *Training corpus SETimes.SR 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1200.

[Blagoeva et al.2012] Diana Blagoeva, Svetla Koeva, and Vladko Murdarov. 2012. *The Bulgarian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at http://www.meta-net.eu/whitepapers.

[Božović et al.2018a] Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018a. *English-Montenegrin parallel corpus of subtitles Opus-MontenegrinSubs 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1176.

[Božović et al.2018b] Petar Božović, Tomaž Erjavec, Jörg Tiedemann, Nikola Ljubešić, and Vojko Gorjanc. 2018b. Opus-montenegrinsubs 1.0: First electronic corpus of the montenegrin language. In Darja Fišer and Andrej Pančur, editors, *Proceedings of the conference on Language Technologies and Digital Humanities 2018*, pages 24–28, Slovenia. Ljubljana University Press.

[Krek2012] Simon Krek. 2012. *The Slovene Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at http://www.meta-net.eu/whitepapers.

[Ljubešić and Klubička2016a] Nikola Ljubešić and Filip Klubička. 2016a. *Croatian web corpus hrWaC 2.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1064.

[Ljubešić and Klubička2016b] Nikola Ljubešić and Filip Klubička. 2016b. *Serbian web corpus srWaC 1.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1063.

[Ljubešić et al.2018a] Nikola Ljubešić, Željko Agić, Filip Klubička, Vuk Batanović, and Tomaž Erjavec. 2018a. *Training corpus hr500k 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1183.

[Ljubešić et al.2018b] Nikola Ljubešić, Tanja Samardžić, Tomaž Erjavec, Darja Fišer, Maja Miličević, and Simon Krek. 2018b. "Kad se mnogo malih složi": Collaborative development of gold resources for Slovene, Croatian and Serbian. In *SlaviCorp 2018 Book of Abstracts*.

[Ljubešić et al.2019a] Nikola Ljubešić, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. 2019a. *Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1241.

[Ljubešić et al.2019b] Nikola Ljubešić, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. 2019b. *Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1240.

[Ljubešić2019a] Nikola Ljubešić. 2019a. *Inflectional lexicon hrLex 1.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1232.

[Ljubešić2019b] Nikola Ljubešić. 2019b. *Inflectional lexicon srLex 1.3*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1233.

[Machálek2020] Tomáš Machálek. 2020. KonText: Advanced and flexible corpus query interface. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France, May. European Language Resources Association.

[Rychlý2007] Pavel Rychlý. 2007. Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

[Simov and Osenova2008] Kiril Simov and Petya Osenova. 2008. Bulgarian Language Resources for Information Technology. In *L. Iomdin and L. Dimitrova (eds.): Lexicographic tools and techniques, Proceedings from MONDILEX 1st Open workshop, 3–4 October, 2008, pp. 60-67*.

[Tadić et al.2012] Marko Tadić, Dunja Brozović-Rončević, and Amir Kapetanović. 2012. *The Croatian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at http://www.meta-net.eu/whitepapers.

[Vitas et al.2012] Duško Vitas, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. 2012. *The Serbian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer. Available online at http://www.meta-net.eu/whitepapers.

sticker2: a Neural Syntax Annotator for Dutch and German

Daniël de Kok and **Neele Falk** and **Tobias Pütz**
Seminar für Sprachwissenschaft
University of Tübingen, Germany
{daniel.de-kok, neele.falk, tobias.puetz}@uni-tuebingen.de

Abstract

In this work, we introduce sticker2, a production-quality, neural syntax annotator for Dutch and German based on deep transformer networks. sticker2 uses multi-task learning to support simultaneous annotation of several syntactic layers (e.g. part-of-speech, morphology, lemmas, and syntactic dependencies). Moreover, sticker2 can finetune pretrained models such as XLM-RoBERTa (Conneau et al., 2019) for state-of-the-art accuracies.

To make use of the deep syntax models tractable for execution environments such as WebLicht (Hinrichs et al., 2010), we apply model distillation (Hinton et al., 2015) to reduce the model's size. Distillation results in models that are roughly 5.2-8.5 times smaller and 2.5-4.4 times faster, with only a small loss of accuracy.

sticker2 is widely available through its integration in WebLicht. Nix derivations and Docker images are made available for advanced users that want to use the models outside WebLicht.

1 Introduction

WebLicht (Hinrichs et al., 2010) is an environment for building and executing natural language processing chains. The primary goal of WebLicht is to provide easy access to a wide range of text processing tools to researchers in the humanities and social sciences. WebLicht has changed and grown considerably since its introduction 10 years ago. Its data exchange format, Text Corpus Format (Heid et al., 2010), has been updated to accommodate additional annotation layers, such as chunking and topological fields. The visualization of annotation layers in WebLicht has been improved considerably, and fine-grained search of annotations is now possible (Chernov et al., 2017). The number of annotation services, along with the types of annotations they provide and the languages that they support, has also grown steadily. Currently, WebLicht can apply syntactic analysis such as part-of-speech tagging (17 tools), lemmatization (11 tools) and dependency parsing (11 tools) for 46 languages. WebLicht also provides 6 different named entity recognition systems that work on four different languages.

A recent focus in WebLicht has been the addition of annotation tools based on neural network models. Neural network models provide state-of-the-art results for most natural language processing tasks, outperforming prior non-neural models. In this work, we introduce one such tool, sticker2. sticker2 is a high-accuracy, production-focused neural syntax annotator for Dutch and German, providing part-of-speech tag, lemma, morphology, dependency, and topological field (German) annotations. Since sticker2 combines existing ideas from the literature, this paper will focus on the design choices that were made to make sticker2 ready for production use. In Section 2 we will discuss the architecture of sticker2. The models for Dutch and German are discussed in Section 3. Finally, we will discuss the integration of sticker2 into WebLicht in Section 4.

2 sticker2

2.1 Architecture

At its core, sticker2 is a *sequence labeler* – it assigns, per task, a label to each token in a sentence. In order to do so, sticker2 first splits tokens into smaller pieces (Wu et al., 2016). For instance, the Dutch words *handelsorganisatie* 'trade organization' and *klopt* 'is correct/pulsates' are split into *handel-s-organisatie* and *klopt-t* respectively. This splitting of compounds and/or morphemes enable the use of relatively small vocabularies. Each such piece is represented as a mathematical vector that captures similarities between pieces. These vectors are fed to a transformer network (Vaswani et al., 2017). A tranformer network consists of multiple layers, where in each layer the representation of a piece is updated by attending to the representations of a sentence's pieces in the previous layer. This results in contextualized representations of each piece. For example, when *klop-t* co-occurs with *hart* 'heart', its representation could be specialized for its 'pulsation' sense. These contextualized representations are then weighted differently for each specific task (Peters et al., 2018; Kondratyuk and Straka, 2019) to extract the information that pertains to that task. Finally, distributions over possible labels are obtained by applying a classification layer to the task-specific representations. The model is trained end-to-end and simultaneously on all tasks (*multi-task training*). This allows the model to learn joint contextual representations for all tasks in the transformer.

Since such transformer networks typically require a larger amount of data to train than what is available in supervised training sets, sticker2 supports finetuning of pretrained models (Devlin et al., 2019; Conneau et al., 2019). These models are usually trained on a general task for which no supervised data is required, such as predicting randomly-masked words.

Besides basic sequence labeling, sticker2 also supports two forms of structural predictions: lemmatization and dependency parsing. In lemmatization, sticker2 predicts edit trees (Chrupała, 2008), which are applied to tokens after prediction to infer their lemmas. For dependency parsing, we use the dependency encoding scheme proposed by Spoustová and Spousta (2010). In this scheme, every token is annotated with a triple that contains: (1) the dependency relation of the token to its head, (2) the part-of-speech tag of the head; and (3) the relative position of the head in terms of its part-of-speech tags. For instance, the tag `nmod/noun/-2` means that a token should be attached with the *nmod* relation to the second preceding noun. After tagging, such tags are decoded to construct the dependency structure of a sentence.

2.2 Amenities for production use

In order to support sticker2 in production setups, we have focused on several aspects: speed, memory use, parallelization, and deployability. Prediction speed and memory use are, by and large, dominated by the size of the transformer network. In particular, prediction speed is dominated by the number of hidden layers and the hidden layer size, whereas memory use is dominated by the hidden layer size and the number of word or sentence piece embeddings.

To reduce the number of hidden layers and the hidden layer size, sticker2 supports model distillation (Hinton et al., 2015). Model distillation trains a new, smaller student model on the predictions of a larger teacher model on a large, unsupervised training set. After distillation, the student model is fine-tuned on the supervised training set. As we will show in Section 3, model distillation will result in drastically smaller and faster models at a marginal loss in accuracy. In addition to performing model distillation, we reduce the size of the models further by using smaller vocabularies in the student models.

We have implemented sticker2 in the Rust programming language.[1] The transformer models are also implemented in Rust using the linear algebra and back-propagation primitives that are made available through `libtorch` (Paszke et al., 2019). Rust programs compile to a single binary that can be deployed without requiring an additional language runtime or runtime packages. Rust's strong safety guarantees allow us to run sticker2 virtually lock-free in a single process. This makes it possible to deploy a single server process that can serve a large number of clients concurrently, while sharing resources such as the model.

[1] `https://www.rust-lang.org/`

3 Dutch and German syntax models

In this section, we describe the Dutch and German syntax models that we make available for sticker2. For both languages, we provide a fine-tuned XLM-RoBERTa base (Conneau et al., 2019) model, as well as smaller distilled models.

Dutch The Dutch model was fine-tuned and evaluated on the *Universal Dependencies* (UD) (Nivre et al., 2016) conversion of the Lassy Small treebank (Van Noord et al., 2013; Bouma and van Noord, 2017). This treebank consists of 65,147 sentences and 1,095,087 tokens. We split a random shuffle of the treebank sentences in 70/10/20% portions to obtain training, development, and held-out sets. We fine-tune the XLM-RoBERTa base model on the *universal part-of-speech tag*, *lemma*, *morphology tag*, and *universal dependency* layers of the treebank.

As discussed in Section 2.2, we use model distillation to create smaller and faster models. Since XLM-RoBERTa uses a large (multi-lingual) vocabulary of 250,000 sentence pieces, we use a smaller vocabulary of 30,000 word pieces for the distilled models.[2] Given the XLM-RoBERTa model, which has 12 layers (*l=12*), 768 hidden units (*h=768*), and 12 attention heads (*hd=12*), we extract the following two models: (1) h = 368, l = 12, hd = 12; and (2) h = 368, l = 6, hd = 12. The model was first distilled on the Lassy Large corpus (Van Noord et al., 2013) minus the sentences of Lassy Small, leading to a corpus of 47.6M sentences and 700M tokens. After distillation, the model is fine-tuned on the training set.

German The German model was trained on the UD conversion of TüBa-D/Z (Telljohann et al., 2005; Çöltekin et al., 2017), which consists of 104,787 sentences and 1,959,474 tokens. We use the same 70/10/20% split as for Dutch. The model is fine-tuned on the same layers as those described for Dutch, with the addition of a topological field layer. We distill models of the same transformer network and vocabulary sizes as for Dutch. The models are distilled on a mixture of Taz newspaper and Wikipedia subsections of the TüBa-D/DP (de Kok and Pütz, 2019) minus the sentences of TüBa-D/Z, consisting of 33.8M sentences and 648.6M tokens.

Results The accuracies of the Dutch and German models can be found in Table 1 and Table 2 respectively. The accuracy of dependency parsing is reported in *labeled attachment score* (LAS), which is the percentage of tokens that have the correct head and dependency relation. The tagging speed was measured in sentences per second on the held-out data on a Core i5-8259U mobile CPU with 4 threads. These results show that we can reduce the size and improve the speed of the models drastically for production use, such as in WebLicht, with relatively small reductions in accuracy.

Model	POS	Lemma	Morph	LAS	Size (MB)	Sent/sec
XLM-RoBERTa	98.89	99.04	98.87	93.13	1003	44
l = 12, h = 384, hd = 12	98.81	99.05	98.82	93.35	194	112
l = 6, h = 384, hd = 12	98.80	99.01	98.78	93.09	127	194

Table 1: Performance of XLM-RoBERTa and distilled models on the Lassy Small held-out data set.

Model	POS	Lemma	Morph	TF	LAS	Size (MiB)	Sent/sec
XLM-RoBERTa	99.24	99.33	98.35	98.14	95.59	1107	35
l = 12, h = 384, hd = 12	99.20	99.31	98.33	98.14	95.77	199	88
l = 6, h = 384, hd = 12	99.18	99.28	98.27	98.03	95.33	131	147

Table 2: Performance of XLM-RoBERTa and distilled models on the TüBa-D/Z held-out set.

[2]The reduction from 250,000 to 30,000 word pieces reduces the size of the word piece embeddings from 366MiB to 44MiB in the distilled models.

4 Integration into WebLicht

As shown in the previous section, the distilled models are much smaller and faster and can therefore be easily put into production. The WebLicht web services hosted in Tübingen are tested and deployed in production as *Docker* images. sticker2, having been integrated into WebLicht, also runs within its own isolated environment provided by the Docker platform. Several Docker containers are working together to offer the sticker2 service: One for each sticker2 worker and another one for a central service that is responsible for data conversion and communication. The central service, acting as a front-end to client requests, connects to the sticker2 workers, retrieves the processed data and converts it into the Text Corpus Format. This way the users can call the sticker2 web service via the WebLicht user interface and retrieve the syntactic annotations for their own data in a simple way.

5 Conclusion

In this work, we have introduced the sticker2 syntax annotation tool, as well as models for Dutch and German. We have shown that by using techniques such as model distillation, models can be made small and fast enough for high-accuracy annotation of large text corpora. Finally, we have described how sticker2 has been integrated into WebLicht, making it available as part of the CLARIN infrastructure.

sticker2 is available in the WebLicht interface[3] as *Sticker2-Dutch* and *Sticker2-German* or as a predefined easy chain. Advanced users can also use Nix derivations[4] or Docker images published through Docker Hub.[5]

In the near future, we hope to add support for additional languages, as large UD treebanks for those languages become available.

6 Acknowledgements

We would like to thank Erhard Hinrichs and Patricia Fischer for their feedback on this work. The financial support of the research reported in this paper is provided by the German Federal Ministry of Education and Research (BMBF), the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK) as part of CLARIN-D and by the German Research Foundation (DFG) as part of the Collaborative Research Center 'The Construction of Meaning' (SFB 833), project A3.

References

Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: the automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, 5.

Alexandr Chernov, Erhard W. Hinrichs, and Marie Hinrichs. 2017. Search your own treebank. In Markus Dickinson, Jan Hajic, Sandra Kübler, and Adam Przepiórkowski, editors, *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15), Bloomington, IN, USA, January 20-21, 2017*, volume 1779 of *CEUR Workshop Proceedings*, pages 25–34. CEUR-WS.org.

Grzegorz Chrupała. 2008. *Towards a machine-learning architecture for lexical functional grammar parsing*. Ph.D. thesis, Dublin City University.

Çağrı Çöltekin, Ben Campbell, Erhard Hinrichs, and Heike Telljohann. 2017. Converting the TüBa-D/Z treebank of German to Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 27–37, Gothenburg, Sweden, May.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Daniël de Kok and Sebastian Pütz. 2019. *Stylebook for the Tübingen treebank of dependency-parsed German (TüBa-D/DP)*. Seminar fur Sprachwissenschaft, Universitat Tübingen, Tübingen, Germany.

[3] `https://weblicht.sfs.uni-tuebingen.de/weblicht/`
[4] `https://github.com/stickeritis/nix-packages`
[5] `https://hub.docker.com/repository/docker/danieldk/sticker2`

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard Hinrichs. 2010. A corpus representation format for linguistic web services: The D-SPIN text corpus format and its relationship with ISO standards. In *Proceedings of LREC 2010*, Valletta, Malta, May.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden, July.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of EMNLP-IJCNLP 2019*, pages 2779–2795, Hong Kong, China, November.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June.

Drahomíra Spoustová and Miroslav Spousta. 2010. Dependency parsing as a sequence labeling task. *The Prague Bulletin of Mathematical Linguistics*, 94:7–14.

Heike Telljohann, Erhard W Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2005. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). In *Seminar für Sprachwissenschaft, Universität Tübingen, Germany*.

Gertjan Van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer Van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In *Essential speech and language technology for Dutch*, pages 147–164. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Exploring and Visualizing Wordnet Data with GermaNet Rover

Marie Hinrichs **Richard Lawrence** **Erhard Hinrichs**

University of Tübingen, Germany
{marie.hinrichs, richard.lawrence, erhard.hinrichs}@uni-tuebingen.de

Abstract

This paper introduces GermaNet Rover, a new web application for exploring and visualizing GermaNet, the German wordnet. Rover provides semantic relatedness calculations between concept pairs using six algorithms, and allows regular expression or edit distance lookup in combination with other search constraint options. Visualizations include a concept's position in the hypernym graph and the shortest path between concepts. Rover provides easy access to these features and is available as a CLARIN resource.

1 Introduction

GermaNet[1] is a lexical-semantic network (a *wordnet*) for the German language. Wordnets have been developed for many languages, often modeled on the Princeton WordNet® for English (Fellbaum, 1998). A number of wordnets are available as CLARIN[2] resources – for example plWordNet (Maziarz et al., 2016), Estonian Wordnet (Kahusk and Vider, 2005; Pedersen et al., 2013), FinnWordNet (Lindén and Carlson., 2010), Open Dutch Wordnet (Postma et al., 2016), and BulNet (Rizov and Dimitrova, 2016).

There are many web applications for browsing wordnets, several of which include visualizations. But current tools generally lack several features which would be useful for researchers. For example, many tools only include simple searching capabilities, and lack a means of specifying non-exact, "fuzzy" search terms. Many tools also lack a means of comparing the concepts in the wordnet, although wordnet-based measures of semantic similarity have been available for some time. To the best of our knowledge, plWordNet offers the only web application which presents semantically similar words for a given concept, and only one similarity score is presented.

GermaNet Rover[3] is a new web application for exploring and visualizing GermaNet. Rover provides advanced search features, including searching by regular expression or edit distance, as well as a variety of other search constraints. It allows comparing concepts in the wordnet via six different measures of semantic relatedness, which can be viewed in the web interface for individual pairs of concepts, or processed in batch via file upload. Rover also offers visualizations of the network structure on which these measures are based, including of a concept's position in the wordnet and of the shortest paths between two concepts. Rover provides user-friendly access to these features and is available as a CLARIN resource.

2 Background

A wordnet is a data set representing a network of semantic relationships in a language. In a wordnet, words or *lexical units* are grouped into sets of synonyms called *synsets*. Each synset represents a single concept of the language, which may be variously expressed by the lexical units it contains. Relations

[1] https://uni-tuebingen.de/en/142806
[2] https://www.clarin.eu/
[3] https://weblicht.sfs.uni-tuebingen.de/rover/

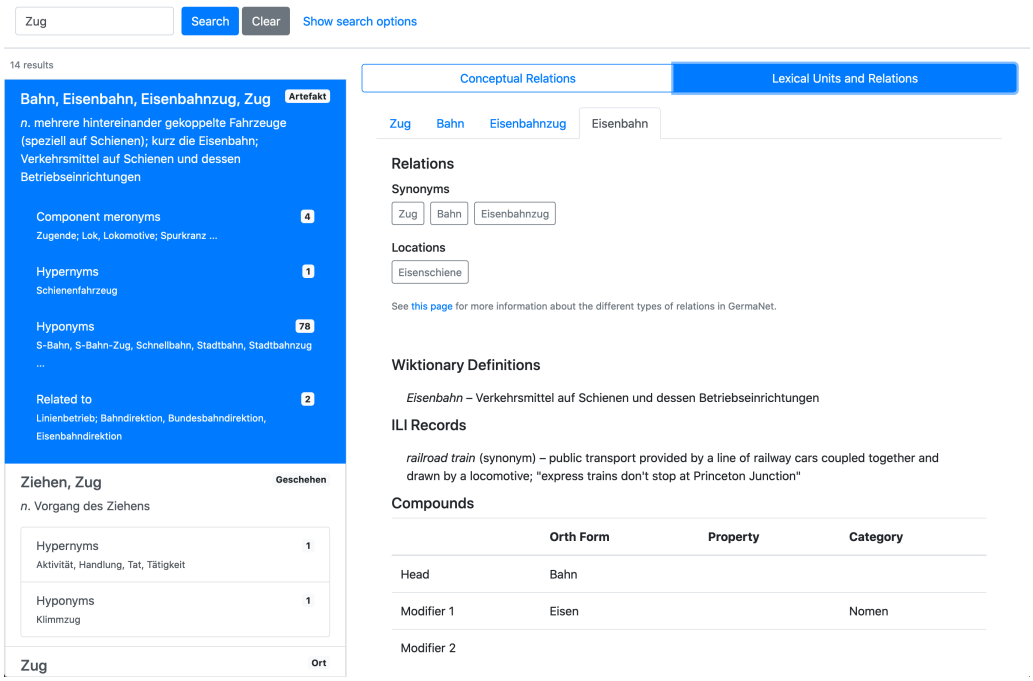


Figure 1: The Synset Search interface, with results of a search for *Zug* "train". Summaries of synsets in the search results appear on the left. Details about the selected synset appear on the right.

between synsets, called *conceptual relations*, are a wordnet's primary representation of semantic relationships. GermaNet also contains *lexical* relations, which hold between individual lexical units, rather than synsets.

There are several types of conceptual relations in GermaNet. The primary relation is *hypernymy*, an asymmetric predication relation ('X is Y but Y is not necessarily X'). For example, the synset containing *Haustier* "pet" has as hypernym the synset containing *Tier* "animal". Other conceptual relations in GermaNet include *causation*, *entailment*, and *meronomy*, the latter of which represents whole-part relationships (e.g. *Pedal* "pedal" is part of a *Fahrrad* "bike").

The hypernymy relation is uniquely important because every synset is guaranteed to have at least one hypernym, except for the artificial root synset. The synsets thus form a connected directed graph with edges given by the hypernymy relation. This graph has a mostly-hierarchical structure, in which more general concepts may be thought of as being higher in the hierarchy, with more specific concepts falling under them. This graph structure is the basis for several measures in the literature of semantic relatedness between concepts.

3 GermaNet Rover

GermaNet Rover is a web application that enables researchers to explore the semantic relationships and other data in GermaNet. Rover currently has two main features: Synset Search and Semantic Relatedness. Both features also provide relevant visualizations of the local structure of GermaNet. We describe these features in more detail below.

3.1 Synset Search

The Synset Search feature (see Figure 1) allows searching for synsets in GermaNet, browsing the relationships between synsets, and viewing detailed data about individual synsets, including a visualization of its place in the hypernym relation.

Searches match a search term against all the synsets in GermaNet. The search space can be further

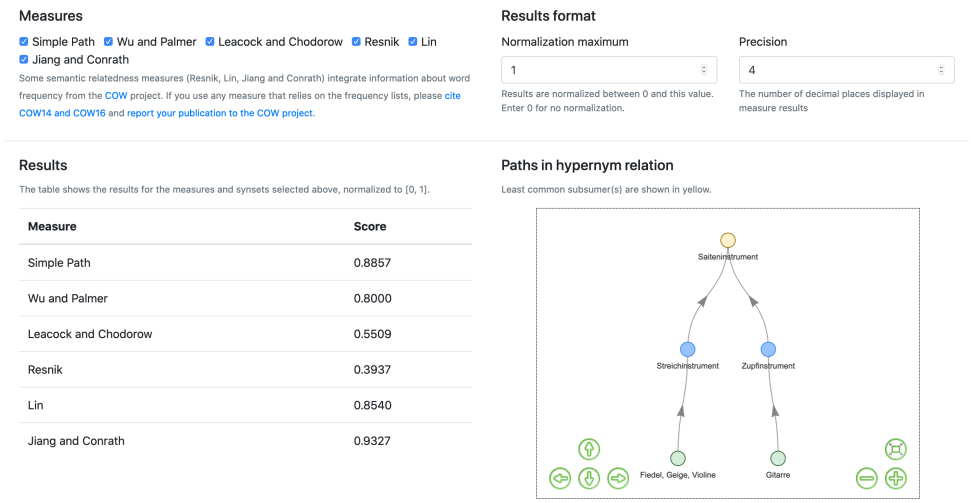


Figure 2: The Visualize Relatedness interface in the Semantic Relatedness feature, comparing *Gitarre* "guitar" and *Fiedel, Geige, Violine* "violin". The results table displays relatedness scores for each selected measure. The network diagram displays the shortest path between the two synsets via their least common subsumer.

constrained to particular grammatical categories (e.g. adjectives), semantic classes (e.g. *Bewegung* "motion"), or orthographic variants (including forms used prior to the orthography reform of 1996, and variant forms of both the current and old forms). Fuzzy matching of the search term is supported through three different methods: by ignoring capitalization in the search term, by specifying a numerical edit distance from the search term, or by treating the search term as a regular expression. The search parameters are matched against the individual lexical units in a synset; any synset containing at least one lexical unit matching the search parameters will be returned as a result. For each synset in the result set of a search, a brief summary appears with the synset's word category and class, any definitions or paraphrases associated with the synset, and the type and number of the synset's conceptual relations.

Selecting one of these summaries displays more detailed information about the synset, separated into two tabs. The Conceptual Relations tab offers a detailed presentation of all of the synset's conceptual relations, organized by type. For each related synset, the lexical units of that synset are displayed as a series of buttons which navigate to the related synset, which allows progressively exploring the network via its relations. The Conceptual Relations tab also displays a network diagram which visualizes the local structure of the hypernymy relation between the selected synset and the artificial root synset.

The Lexical Units and Relations tab is further broken down into one detail tab for each lexical unit in the selected synset. These detail tabs display any information present in GermaNet about these individual lexical units, including their relations to other lexical units, example sentences with associated frame types, and associated records in other data sets.

3.2 Semantic Relatedness

The Semantic Relatedness feature allows calculating the relatedness between pairs of synsets using several different algorithms. The measures currently implemented are derived from Wu and Palmer (1994), Jiang and Conrath (1997), Leacock and Chodorow (1998), Lin (1998), and Resnik (1999). An additional *Simple Path* measure calculates the length of the path between the two synsets via the hypernymy relation, relative to the length of the longest such path. Because the different measures have different minimum and maximum values, the calculations are normalized to a common interval, so that the results can be more easily compared.

The Semantic Relatedness feature provides two different interfaces to work with these measures: Visualize Relatedness and Batch Processing. On the Visualize Relatedness tab (see Figure 2), two specific

synsets are selected using two independent search inputs, similar to those in the Synset Search feature. The semantic relatedness of the two selected synsets is then calculated according to the implemented measures, and the results are displayed in a table. The display of the results table can be controlled by selecting different measures and adjusting the normalization interval.

An important notion underlying each of the implemented measures is the *least common subsumers* of the selected synsets. These are the two synsets' nearest common ancestors in the hypernymy relation. The Visualize Relatedness tab also displays a network diagram which visualizes the shortest paths between the two selected synsets via their least common subsumers. Both the results table and the network diagram are updated automatically whenever different synsets or measures are selected, allowing quick exploration of how the different measures behave in relation to the structure of the hypernym graph.

The Batch Processing tab can be used to calculate the relatedness of many synsets at once by uploading a file containing word pairs. The batch processor searches for synsets containing the words in each pair, and then compares each pair of synsets containing the two words. (Thus, a single word pair can generate multiple synset comparisons.) A configuration section in the input file allows specifying the same options as are available on the Visualize Relatedness tab, including constraints for the searches, a selection of individual measures, and the normalization interval for the calculations. Results are returned in a tabular format that can easily be processed with a spreadsheet program or custom scripts.

3.3 Advances in Rover

Previous work had also enabled exploring and visualizing the data in GermaNet (Finthammer and Cramer, 2008; Henrich and Hinrichs, 2010). Rover builds both on that work and on several other internal projects, offering new advantages and opportunities for researchers working with wordnet data.

Unlike previous tools for GermaNet, Rover is structured as a web application, with a browser-based interface. This eliminates the need for users who just want to explore the data set to download or install special software, and allows the data and backend web server to be hosted on CLARIN infrastructure.

The user interface of Rover has been implemented using a new open source library called germanet-common[4], which provides abstractions for querying and displaying different types of wordnet data from a web server via a JSON API. Due to the design of this library, the Rover web interface would be relatively easy to port to other wordnets' infrastructure, and could provide a basis for exploring and visualizing data in other wordnets. The library also makes it easy to build other kinds of web interfaces for wordnets, beyond the features available in Rover.

The implementation of Rover has also brought new developments in the GermaNet Java API[5], the central library used by Rover's web server. This library implements the fuzzy matching capabilities in the Synset Search feature and the various measures in the Semantic Relatedness feature. Since it is released annually alongside the data set, these features are not limited to the Rover interface and will also be available to researchers working with the GermaNet data via this API.

4 An Application: Morphological Productivity

Beyond simple uses as a dictionary or thesaurus, Rover also supports more advanced linguistic research. This section presents an example of a research question that can quickly be studied with Rover.

Morphological productivity concerns the extent to which a morpheme can productively form compounds. The regular expression and grammatical category support in the Synset Search feature provides a way to quickly get a sense of the relative productivity of a given morpheme. For example, how productive is the adjective *reich* "rich" in comparison to its antonym *arm* "poor"?

To answer this question, one might start by searching with the regular expressions `.+reich` and `.+arm`. The leading `.+` in these search terms matches any non-empty string of characters, so they will match adjective compounds like *erfolgreich* "successful" and *abgasarm* "low-emission". The need to refine the searches will quickly become obvious from the results: one will want to exclude nouns like *Frankreich* "France" by limiting the results to adjectives, as well as adjectives that are compounds of

[4] `https://github.com/Germanet-sfs/germanet-common`
[5] `https://github.com/Germanet-sfs/GermaNetApi/`

warm rather than *arm* by adjusting the search term to `[^w]+arm`. With these refinements, one learns that *reich* is about three times as productive as *arm* in adjective compounds (138 vs. 46 results). Rover is designed to support this kind of interactive exploration and refinement, which can help researchers in their study of the German language.

5 Future work

Rover is still being actively developed, and several features are planned for future work. We are investigating the possibility of including additional measures, based on word embeddings, in the Semantic Relatedness feature. We also hope to further improve the network visualizations. Feedback from CLARIN researchers and beta-testers will inform this work.

Acknowledgements

The research described in this paper was carried out as part of the CLARIN-D project, which is funded by the German Federal Ministry of Education and Research (BMBF).

References

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Marc Finthammer and Irene M. Cramer. 2008. Exploring and Navigating: Tools for GermaNet. In *Proceedings of the 6th Language Resources and Evaluation Conference*.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT: A graphical tool for germanet development. In *Proceedings of the ACL 2010 System Demonstrations*, pages 19–24. Association for Computational Linguistics.

Jay J. Jiang and David W. Conrath. 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing.

Neeme Kahusk and Kadri Vider. 2005. TEKsaurus—the Estonian WordNet online. In *The Second Baltic Conference on Human Language Technologies: Proceedings.*, pages 273–278.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In *WordNet: An Electronic Lexical Database*. MIT Press.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304, San Francisco, CA, USA.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet — wordnet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140. In Swedish with an English abstract.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 2259–2268, Osaka, Japan.

Bolette S. Pedersen, Lars Borin, Markus Forsberg, Neeme Kahusk, Krister Lindén, Jyrki Niemi4 Niklas Nisbeth, Lars Nygaard, Heili Orav, Eirikur Rögnvaldsson, and Mitchel Seaton. 2013. Nordic and Baltic wordnets aligned and compared through "WordTies". In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pages 147–162.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch Wordnet. In *Proceedings of the Eighth Global Wordnet Conference*, Bucharest, Romania.

P. Resnik. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, July.

Borislav Rizov and Tsvetana Dimitrova. 2016. Hydra for Web: A Browser for Easy Access to Wordnets. In *Proceedings of the Eighth Global Wordnet Conference*, Bucharest, Romania.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Las Cruces, New Mexico, June. Association for Computational Linguistics.

Named Entity Recognition for Distant Reading in ELTeC

Francesca Frontini
Praxiling CNRS
Université Paul-Valéry Montpellier 3
francesca.frontini@univ-montp3.fr

Carmen Brando
CRH,
EHESS, Paris
carmen.brando@ehess.fr

Joanna Byszuk
Institute of Polish Language
Polish Academy of Sciences
joanna.byszuk@ijp.pan.pl

Ioana Galleron
LaTTiCe CNRS
Université Sorbonne Nouvelle - Paris 3
ioana.galleron@sorbonne-nouvelle.fr

Diana Santos
Linguateca &
University of Oslo
d.s.m.santos@ilos.uio.no

Ranka Stanković
University of Belgrade
ranka.stankovic@rgf.bg.ac.rs

Abstract

The "Distant Reading for European Literary History" COST Action, which started in 2017, has among its main objectives the creation of an open source, multilingual European Literary Text Collection (ELTeC). In this paper we present the work carried out to manually annotate a selection of the ELTeC collection for Named Entities, as well as to evaluate existing NER tools as to their capacity to reproduce such annotation. In the final paragraph, points of contact between this initiative and CLARIN are discussed.

1 Introduction

The Distant Reading for European Literary History (COST Action CA16204) kicked off in 2017 with the goal of using computational methods of analysis for large collections of literary texts. The objective is to establish shared practices in the application of innovative computational methods, while at the same time reflecting on the impact that such methods have on our capacity to raise and answer new questions about literary history and theory. More details are to be found on the Action's website[1].

One of the most ambitious deliverables of this COST Action is the creation of a multilingual open source collection, named European Literary Text Collection (ELTeC). In its final version, the corpus will contain at least 10 linguistically annotated subcollections of 100 novels per language (1840-1920[2]), that is at least 1,000 full-text novels. To make subcollections representing particular languages as comparable as possible, the novels are selected to represent a) various types of novels as to their length: short stories, epic novels, b) five twenty-year time periods within the examined time span, c) text of various levels of canonicity, as judged by the number of reprints, and d) as equal as possible ratio of female and male authors. As of now, a first version of the ELTeC corpus has been collected and published with a light TEI encoding[3], with more language collections to be included in the further releases throughout the duration of the Action. Since obtaining more works for some language collections is possible, the Action also plans the publication of the extended ELTeC (including more texts per language or ones published slightly before the assumed time span) estimated to take the total number of full-text novels to at least 2,500.

A case study on Named Entity annotation was carried out in the Working Group 2 "Methods and Tools" (WG2), with an aim of establishing common annotation guidelines – which are specifically oriented to answering a set of scholarly research questions identified by the Action participants – and testing a selection of NER tools, in order to assess their capacity of automatic reproduction of such annotation, a task crucial for annotating corpora of this and bigger sizes.

In this abstract we shall introduce the desiderata for the NE annotation, the current state of the multilingual NE corpus and of the annotation, describe the evaluation framework and provide preliminary results for a selection of NE systems. Finally, we will discuss the relevance of these activities with respect

[1] https://www.distant-reading.net

[2] Chronological limits are due to constraints related to copyright and availability of quality full texts.

[3] https://distantreading.github.io/ELTeC/index.html

to the CLARIN community as well as to the CLARIN services and tools, with some ideas for possible collaboration.

2 Developing the NE layer of the ELTeC corpus

2.1 Desiderata and annotation set

NER is a well known task in NLP, and there are several sets of guidelines for performing NER annotation (see among others (Nadeau and Sekine, 2007; Chinchor, 1998; LDC, 2008; Santos et al., 2006; Zupan et al., 2017)), establishing a number of categories and rules for creating annotated corpora, so that automatic algorithms can try to replicate such annotation. However, NER is a preliminary module for other tasks, such as information extraction. The philosophy guiding the definition of the best known NE guidelines is mostly geared towards very specific scenarios, such as extraction of events in news or scientific texts. Their application to literary texts is therefore not straightforward, despite current attempts at producing new annotation guidelines and datasets (Bamman et al., 2019).

Within the context of the COST action, WG3[4] came up with a set of research questions in order to help with the definition of a more targeted set of annotations. A first set of desiderata emanates from the idea that a novel is an epic set in the private space of a bourgeois home, something which demands researchers to be able to detect indicators of social structure and roles, such as honorifics, names of professions, etc. Another set of research topics touches upon questions about identity, otherness, but also the distinction between urban and rural spaces, which require the annotation of demonyms, as well as a higher granularity in the annotation of toponyms, to facilitate detecting different types of locations (cities vs villages and countryside)[5]. Finally, questions about cultural references, role models and cosmopolitanism can only be answered if references to works of art, authors, folklore and periodical publications are detected.

Such considerations led to the inclusion of categories such as demonyms (DEMO), professions and titles (ROLE), works of art (WORK) in the tagset, alongside more canonical categories, such as person names (PERS), places (LOC), events (EVENT), organisations (ORG) (see Table 1). However, this annotation was later simplified to avoid nested annotations and overlap.

	DEMO	EVENT	LOC	ORG	OTHER	PERS	ROLE	WORK
cze	163	5	275	0	0	1150	454	0
deu	66	2	323	12	0	973	458	4
eng	56	7	198	37	0	1184	203	25
fra	77	3	262	22	128	900	244	18
hun	29	7	152	20	0	1091	367	7
nor	4	8	83	25	3	990	201	10
por1	17	9	351	19	0	940	490	54
por2	34	1	256	30	7	1059	347	7
slv	133	54	336	37	0	1230	620	2
srp	121	18	185	11	0	985	301	4
	700	114	2425	213	138	10514	3685	131

Table 1: Data on the manually NE-annotated corpus.

2.2 Current state of the corpus

The NE annotation of the corpus is part of the plan for the so called level 2 annotation, which will also include morpho-syntactic and direct speech annotation. At this moment, WG2 carries out the NE annotation for a subset of languages: Czech (cze), German (deu), English (eng), French (fra), Hungarian (hun), Norwegian (nor), Portuguese[6] (por1, por2), Slovene (slv), and Serbian (srp). For each language

[4]Working Group 3, dedicated to Literary Theory and History `https://www.distant-reading.net/wg-3/`

[5]At this stage Entity Linking is not envisaged for our corpus, and in any case it would not easily solve the problem of distinguishing between types of populated places for past or fictional locations in novels.

[6]For Portuguese, two sets of novel excerpts are available: por1 comprises canonical novels, in modern ortography, while por2 was created from novels with old ortography, non-canonical.

involved, a sample collection was prepared from the novels already available in ELTeC. Every language-specific collection contains 20 files, each of which is composed of excerpts from one novel. Each file is made up of five ~400-word passages randomly selected from the novel.

The annotation teams worked on the same set of guidelines, using the BRAT[7] annotation tool, but of course some adaptations were necessary due to differences between languages. When two persons worked on the same collection, we performed cross-checking. The latest version of the annotated NE corpus is available online, together with the latest version of the annotation guidelines[8]. For more information we refer also a recent paper (Stankovic et al., 2019) presented at the DH Budapest 2019 conference. The results of the latest round of annotations (May 2019) are presented in Table 1.

2.3 Testing automatic NER

Another important activity of WG2 is testing existing tools to assess their capacity to reproduce the envisaged annotation. At this stage testing is performed without previous domain adaptation, and with a preference for tools that are easy to install and use. The rationale behind this choice is to evaluate whether literary scholars without advanced technical skills will be able to use existing NLP technologies in their research.

In this study, we focused on four collections only, in English, French, Portuguese and Serbian, these being the languages the authors of this paper are most familiar with. This first evaluation round allowed us to develop a common evaluation set up. For each collection, we tested two tools: one common for all (spaCy[9]), and another one language specific (Stanford-NER for English[10], SEM for French[11], PALAVRAS-NER for Portuguese (Bick, 2006) and SrpNER for Serbian). BRAT outputs were compared to annotations produced by these tools by using a shared evaluation script, thus guaranteeing the consistency with the agreed upon strategy for identifying hits and misses in terms of entity detection.

In this first round the evaluation of string detection was strict (segments must match exactly); but we are planning to perform a second round with relaxed evaluation. Only the PERS and LOC tags were mapped to similar categories of NER tools, and evaluated. The performance of the tools was evaluated separately for each tag.

2.4 NER results and analysis

Current evaluation results are presented in Table 2. These preliminary results show the difficulty of the task of NE-annotation of literary novels. A strict evaluation of detection is often penalising for PERS, because of honorifics which we chose to include in our annotation, and is further complicated by the fact that the XML annotated input was processed as such by tools which often expect plain text[12]. In most cases, LOC seems to be less problematic for the pre-trained models. We follow with some remarks for each specific language.

Portuguese: It was expected that off-the-shelf Named Entity recognisers that were developed for modern Portuguese would perform significantly worse in the *por2* collection. This was confirmed for PALAVRAS-NER, which showed lower performance for *por2*, but not for spaCy. This NER system had considerably lower results than PALAVRAS-NER, but no significant performance drop between the two sets (in fact, for PERS it fares even better for *por2*). We believe this is because only the easy cases are catered for by spaCy, and those cases do not depend too much on ortography.

English: In spaCy, tokenization issues due to TEI XML tags included in tokens account for 16% of PERS and 38% for LOC errors. It also misses a lot of PERS (21%) due to undetected honorifics (Mr/Mrs/Miss). Stanford NER has better precision for LOC than spaCy and has no problems with TEI XML tags. For both models, some additional training and fine tuning would be needed for better performances.

[7] `https://brat.nlplab.org/` see also (Stenetorp et al., 2012).

[8] `http://brat.jerteh.rs/#/eltec-simplified/`

[9] We used the out-of-the-box model for all languages except for Serbian, for which none was available.

[10] `https://nlp.stanford.edu/software/CRF-NER.shtml`

[11] `https://www.lattice.cnrs.fr/sites/itellier/SEM.html`

[12] This evaluation scenario is realistic in the context of Digital Literary Studies, where digital editions with a minimal TEI encoding are to be further enriched using NLP tools, to be then analysed by literary scholars.

French: The manual corpus contains many PERS and fewer LOC annotations. However, spaCy-fra annotates too many LOC hence the low precision for this category, and SEM-fra annotates too few PERS, hence the low recall for this category. In general, there are many detection issues in particular with entities including determiners. Despite the better performance, spaCy-fra has an odd behaviour due to parsing, tokenising, presence of XML tags, capital letters in the beginning of sentences, and it recognises entities composed of more than 4 tokens, which are actually rare.
Serbian: SrpNER is a rule based system and the best NER tool for Serbian. 11% of missing annotations are related to PERS multiword units with honorifics "g./gđa." (Mr/Mrs). For LOC, missing annotations are celestial bodies, names of the streets and facilites. With the Serbian spaCy model, about 7% of errors are related to TEI XML tags, much fewer than with the English model, as the training set included TEI annotated text. Further improvements are foreseen.

	Cat	Correct	Missing	Spurious	Precision	Recall	Excess
SEM-fra	LOC	73	112	84	0.465	0.395	0.535
	PERS	82	512	115	0.416	0.138	0.584
SPACY-fra	LOC	103	78	468	0.180	0.569	0.820
	PERS	329	194	297	0.526	0.629	0.474
PALAVRAS-por1	LOC	223	63	44	0.835	0.780	0.165
	PERS	816	90	86	0.905	0.901	0.095
SPACY-por1	LOC	225	84	440	0.338	0.728	0.662
	PERS	465	256	374	0.554	0.645	0.446
PALAVRAS-por2	LOC	151	67	91	0.624	0.693	0.376
	PERS	857	133	285	0.750	0.866	0.250
SPACY-por2	LOC	157	57	396	0.284	0.734	0.716
	PERS	569	236	393	0.591	0.707	0.409
Stanford-eng	LOC	98	100	126	0.438	0.495	0.563
	PERS	649	535	399	0.619	0.548	0.381
SPACY-eng	LOC	98	100	170	0.366	0.495	0.634
	PERS	536	648	240	0.691	0.453	0.309
SrpNER-srp	LOC	107	78	19	0.849	0.578	0.151
	PERS	718	267	158	0.820	0.729	0.180
SPACY-srp	LOC	57	128	104	0.354	0.308	0.646
	PERS	553	432	315	0.637	0.561	0.363

Table 2: Results of the strict evaluation, per language and category.

3 Relationship to CLARIN and conclusion

The creation, annotation and publication of the ELTeC corpus falls clearly within the scope of CLARIN activities, and ties with CLARIN are already assured by individual participants in the COST action (such as Tomaž Erjavec, Slovenian National Coordinator). Nevertheless, the presentation of the corpus at the CLARIN conference represents an opportunity to discuss further avenues for collaboration. Following the FAIR principles and CLARIN best practices, we identify the following points.

The final version of the manually checked NE subset of the ELTeC corpus, as described in this paper, represents a new reference resource in the domain of literary NE annotation, especially for the broad spectrum of languages[13] included. As such, its preservation and visibility should be ensured. When it comes to the publication and preservation of the whole richly annotated ELTeC collection once it is completed (a goal set to be achieved with the end of this Action in 2021), the final decision is yet to be made, with current interest in Textgrid[14] and GAMS[15]. Collaboration with CLARIN should ensure the visibility of the resource, by means of metadata harvesting to the CLARIN Virtual Language Observatory and by listing the corpus in initiatives such as the CLARIN Resource Families.

The usability of the ELTeC collection could be further enhanced by making it a part of the Federated Content Search initiative (especially with the further addition of the morpho-syntactic and semantic

[13] Not all languages in ELTeC have been made the object of NE annotation so far, but future additions are envisaged.

[14] Textgrid [`https://textgridrep.org/`], a repository supported by the Göttingen State and University Library (SUB) and the Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen mbH (GWDG); now a CLARIN C-centre.

[15] `http://gams.uni-graz.at/`

layer). This should be facilitated by the fact that our COST action is planning to publish the corpus also on the TXM platform[16], which already implements CQP queries.

Last but not least, the collaboration could touch upon the question of NLP tools and services, their usability and domain adaptation. Given the fact that for most languages only parts of the ELTeC corpora can be manually annotated, what can CLARIN do to achieve quality automatic processing for the rest of the corpus? And more generally, how can this help in creating ad hoc models for similar texts which are not yet included in the ELTeC collection? Right now CLARIN offers easy to use services to process texts, such as the CLARIN Switchboard and Weblicht. However, NER is not yet available for most languages, and when it is, it generally does not support the processing of TEI-XML texts, something which constitutes a major issue in processing as shown in our results. Moreover, the NER models that are currently available are not necessarily adapted for literary studies and do not allow for the annotation of all of the NER categories demanded by ELTeC design.

The presentation will describe the current state of the ELTeC corpus, with a focus on the NE manually annotated subset and on the tests carried out with various NER modules, and will constitute an opportunity to discuss the aforementioned points with experts from CLARIN ERIC and the various national consortia.

References

Bamman, D., Popat, S., and Shen, S. 2019. An Annotated Dataset of Literary Entities. In *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)*, pages 2138–2144, Minneapolis, Minnesota, June. Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-1220.

Bick, E. 2006. Functional Aspects in Portuguese NER. In Vieira, R., Quaresma, P., da Graça Volpes Nunes, M., Mamede, N. J., Oliveira, C., and Dias, M. C., editors, *Computational Processing of the Portuguese Language: 7th International Workshop, PROPOR 2006. Itatiaia, Brazil, May 2006 (PROPOR'2006)*, pages 80–89. Springer Verlag. https://link.springer.com/chapter/10.1007/11751984_9.

Chinchor, N. A. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. https://www.aclweb.org/anthology/M98-1001.

LDC. 2008. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6. Technical report, Linguistic Data Consortium. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf.

Nadeau, D. and Sekine, S. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January. Publisher: John Benjamins Publishing Company, http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002.

Santos, D., Seco, N., Cardoso, N., and Vilela, R. 2006. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odjik, J., and Tapias, D., editors, *Proceedings of LREC 2006 (LREC'2006)*, pages 1986–1991. http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.

Stankovic, R., Santos, D., Frontini, F., Erjavec, T., and Brando, C. 2019. Named Entity Recognition for Distant Reading in Several European Literatures. In *DH Budapest 2019*, Budapest. http://elte-dh.hu/wp-content/uploads/2019/09/DH_BP_2019-Abstract-Booklet.pdf.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics. https://www.aclweb.org/anthology/E12-2021.

Zupan, K., Ljubešić, N., and Erjavec, T. 2017. Annotation guidelines for Slovenian named entities: Janes-NER. Technical report, Jožef Stefan Institute, September. https://www.clarin.si/repository/xmlui/bitstream/handle/11356/1123/SlovenianNER-eng-v1.1.pdf.

[16] `http://textometrie.ens-lyon.fr/`

Towards Semi-Automatic Analysis of Spontaneous Language for Dutch

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

Abstract

This paper presents results of an application (*Sasta*) derived from the CLARIN-developed tool *GrETEL* for the automatic assessment of transcripts of spontaneous Dutch language. The techniques described here, if successful, (1) have important societal impact, (2) are interesting from a scientific point of view, and (3) may benefit the CLARIN infrastructure itself since they enable a derivative program called *CHAMP-NL* (CHAT-iMProver for Dutch) that can improve the quality of the annotations of Dutch data in CHAT-format.

1 Introduction

This paper presents results of an application (*Sasta*) derived from the CLARIN-developed tool *GrETEL* for the automatic assessment of transcripts of spontaneous Dutch language. The techniques described here, if successful, (1) have important societal impact (enabling analysis of spontaneous language in a clinical setting, which is considered important but usually not done because it takes too much effort), (2) are interesting from a scientific point of view (various phenomena get a linguistically interesting treatment), and (3) may benefit the CLARIN infrastructure itself since they enable a derivative program called *CHAMP-NL* (CHAT-iMProver for Dutch) that can improve the quality of the annotations of Dutch data in CHAT-format (CHILDES data, (MacWhinney, 2000)).

2 Analysis of Spontaneous Language

The analysis of spontaneous language is considered an important method for determining the level of language development and for identifying potential language disorders. Crystal et al. (1976) developed the LARSP method for language assessment, remediation and screening.[1] Many researchers developed variants of LARSP for other languages, see e.g. (Ball et al., 2012). Also for the Dutch language various methods have been developed for the analysis of spontaneous language, both for assessment of language development, e.g. GRAMAT (Bol and Kuiken, 1989), TARSP (Schlichting, 2005; Schlichting, 2017), and STAP (van Ierland et al., 2008; Verbeek et al., 2007) as well as for assessment of aphasia, e.g. ASTA (Boxum et al., 2013).

Though analysis of spontaneous language is important, it is very time consuming and requires full concentration, so that it is often not done or done only partially. There is a clear need to investigate whether the process can be automated in full or partially.

This paper reports on initial experiments to partially automate the grammatical annotation stage of this process, with the goal to gain efficiency and possibly also to increase the quality of the annotations.

3 GrETEL

GrETEL (Augustinus et al., 2012) is an application to query treebanks. It makes existing manually verified treebanks for Dutch such as *LASSY-Small* for written Dutch (van Noord et al., 2013) and the *Spoken Dutch Corpus* (Oostdijk et al., 2002) available for search. The syntactic structures inside the treebanks

[1] LARSP= Language Assessment, Remediation and Screening Procedure.

are encoded in XML. GrETEL offers XPath to search in these syntactic structures for words, grammatical properties and constructions. In addition, it offers query-by-example facilities.

Version 4 of GrETEL (GrETEL4, (Odijk et al., 2018)) enables a researcher to upload a text corpus and associated metadata, and have it automatically parsed by the Alpino parser (Bouma et al., 2001), after which the resulting treebank is made available for search. It also offers various ways of analysing the search results, for data and metadata combined.

The text corpus upload functionality also makes it possible to upload a transcript of a spontaneous language session and to analyse it for grammatical properties. We describe experiments with this in section 5.

4 Related Work

To our knowledge, (Bishop, 1984) was the first to propose and partially implement an automation of LARSP (for English). (Long et al., 1996 2000) developed a different system for English. For French, F-LARSP was automated by (Parisse et al., 2012), though it could deal reliably only with inflectional properties and the lower stages of development (stages I-III but not stages IV and V). To our knowledge, no attempts have been made before to automate any of the spontaneous language assessment methods for Dutch. However, there has been work on automating the determination of readability, e.g. with the tools *T-Scan*[2] and *Lint* (Pander Maat and Dekker, 2016; Pander Maat, 2017). Though these are different applications applied to a different domain (prepared written texts) and with different purposes (readability assessment), many of the underlying technologies are shared. For example, T-Scan also uses the Alpino parser. Researchers of the University of Groningen developed a Syntactic Profiler of Dutch (SPOD[3]), as part of the treebank query application PaQu (Odijk et al., 2017). SPOD also targets prepared written texts.

5 Schlichting Appendix Test

In order to assess the potential of GrETEL to automate the TARSP analysis, we experimented on the appendix of (Schlichting, 2005). This appendix is intended for illustrating the TARSP analysis and contains a number of example sentences together with their analysis in the form of annotations. We use the analysis as our reference material. For reasons that will become clear below, we call this the *Bronze* reference. Queries have been written for the TARSP language measures that cover the annotations. Multiple matches can occur in the same utterance, so the queries yield multisets of utterance identifiers. The Bronze reference has also been specified with a multiset of utterance identifiers for each language measure.

We noticed after doing several experiments that GrETEL finds many matches that do not occur in the Bronze reference but that are (in our view) correct. We therefore created a second, improved reference, which we have called the *Silver* reference, which includes the utterance identifiers found by GrETEL that are not in the Bronze reference but considered correct by us. We suspect that the omission of these annotations in the Bronze reference is partially due to human oversight, and partially due to the fact that these data were never created as reference data but rather as illustrative analyses. Though a comparison with a Silver reference probably yields a higher score than a comparison with a truly complete reference (a *Gold* reference), it is a useful way to get an impression of what kind of performance is attainable. Having a Silver reference enables us to do three comparisons: (1) GrETEL v. Bronze reference; (2) GrETEL v. Silver reference; (3) Bronze v. Silver reference.

We use *recall*, *precision* and *F1-score* with their common interpretation as measures. These will be given explicit definitions in the full paper. The results of the experiment have been summarised in Table 1.

The figures that we observe here are promising, though it must of course be noted that the experiment has not been carried out on an independent test set.

[2] `https://webservices-lst.science.ru.nl/tscan`.
[3] `https://paqu.let.rug.nl:8068/spod`

Comparison / Measure	R	P	F1
GrETEL v. Bronze	0.88	0.79	0.83
GrETEL v. Silver	0.89	0.86	0.87
Bronze v. Silver	0.88	0.90	0.89

Table 1: Performance of GrETEL versus a human-created Bronze reference, versus an improved reference called Silver, and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1).

6 SASTA Project

The results described in section 5 were considered promising by ourselves and the Dutch Association of Clinical Linguistics (VKL). For this reason we decided to extend the development, in a project called SASTA (acronym for a Dutch expansion meaning Semi-Automatic Assessment of Spontaneous Language).

In the project we have developed a research prototype application called *Sasta* aimed at clinical linguists that takes as input (1) a transcript to be analysed; and (2) an assessment method to be applied. The application yields as output (1) a standard profiling form in accordance with the assessment method, plus an assessment of the language development stage or the language disorder of the patient; (2) the transcript enriched with annotations. The automatically annotated transcript can be manually adapted and then offered to *Sasta* again for generating a revised profiling form. We support three different assessment methods (TARSP, STAP and ASTA). Each method is defined as a set of queries, special modules that are needed, measures to deal with deviating input, etc. associated to language measures of the method.

In order to develop *Sasta* we have developed *Sastadev*, a piece of software intended for developers that enables input of multiple reference data in multiple formats and compares the output of *Sasta* with the references and provides a detailed analysis of the differences. Many data provided by VKL members and other clinical linguists have been used for training and testing the system.

Sasta and *Sastadev* reuse components of GrETEL (the Alpino parser, the upload functionality, and the query functionality) but apply them differently: GreTEL is optimally suited to apply a single query to a large treebank, while *Sasta* and *Sastadev* are more suited to apply multiple queries to a small treebank.

7 Problems to Be Addressed

There are many problems that the data and the technology pose and that have to be addressed.

First, the transcripts of the spontaneous language sessions contain a large amount of deviations of normal language use. These are partially due to annotation conventions, and partially due to the fact that the children who are still learning the language and patients with aphasia make imperfect utterances.

Conventions for annotating the data had to be made more formal and more detailed, as will be explained in the full paper. We therefore require annotations based on the CHAT-format.

In the full paper we will discuss various kinds of deviations in the transcripts: deviations to indicate the way the word was pronounced, (regional) spoken language variants, wrong pronunciations, false starts, repetitions, incomplete utterances, grammatical errors (e.g. overgeneralisations, wrong determiner, wrong auxiliary, wrong determiner noun combinations, etc.).

Second, the Alpino parser has limitations. It cannot analyse all compounds as compounds, it provides insufficient information on verbless utterances, it provides insufficient information on Verb-first sentences, it sometimes parses an utterance incorrectly, it sometimes analyses an utterance in a way that differs from the reference (but is not incorrect). Alpino does not consider the context, can do very little when semantic restrictions apply, and cannot deal with intonation .

Third, certain items require queries that cannot be expressed in XPath or only with great difficulty, e.g., the TARSP item *6+* which requires 6 or more constituents in a clause.

8 Towards Solutions

Many of the problems identified in section 7 can be addressed and several have already been addressed.

For example, by writing the right queries we can analyse certain adverbs inside phrases as if they occur at a sentential level. For queries that cannot be easily formulated in XPath we enable functions in a full programming language (we use Python). In addition, we allow macros inside XPath queries to make the queries shorter and easier to read and to facilitate reuse.

We developed new modules for normalising orthography, for analysing compounds, for dealing with regional spoken language diminutives ending in *-ie(s)*, for overgeneralised inflectional forms of verbs, and for automatically detecting filled pauses and repetitions, as will be explained in the final paper. We use these to adapt each utterance that Alpino cannot deal with to a variant of this utterance that Alpino can deal with. Some examples have been given in table 2

Original utterance	**Corrected utterance**	**Gloss**
mama mouwe hoog	mama mouwe**n** hoog	mum sleeves high
niet goed uitgekijken	niet goed uitgek**e**ken	not well looked-out
die stukkies	die stuk**j**es	those pieces-DIM
zie-ken-huis	ziekenhuis	hospital

Table 2: Some examples of automatic corrections to improve the performance of Alpino.

We are using data provided by the VKL and by several clinical linguists, all example sentences of (Schlichting, 2005) and Dutch CHILDES data during development. We kept a number of data provided as unseen test data.

9 Recent results

Schlichting	**%**		O v B			O v S			B v S		
Eval Meth	**Corr**	**Exts**	**R**	**P**	**F1**	**R**	**P**	**F1**	**R**	**P**	**F1**
Sastadev	No	No	86.5	80.8	83.6	88.4	89.3	88.8	89.8	97.0	93.3
Sastadev	Yes	No	88.5	81.2	84.7						
Sastadev	No	Yes	88.0	70.1	78.0	89.0	77.1	82.6	86.8	94.4	90.4
Sastadev	Yes	Yes	91.2	70.8	79.7						

Table 3: Performance of Sastadev for the Schlichting Appendix

Table 3 shows the performance of Sastadev for the Schlichting Appendix (O) versus a human-created Bronze reference (B), versus an improved reference called Silver (S), and of the Bronze reference versus the Silver reference in terms of recall (R), precision (P) and F1-score (F1). Results are given for the original version of TARSP, for the original version of TARSP with automatically created corrections, and for the most recent extensions to TARSP (column *exts* in the table) without and with corrections. The corrections applied are certain orthographic normalisations (addition of *n* after a word ending in *e*, separation of incorrectly concatenated words, dehyphenation), normalisation of regional diminutives, normalisation of overgeneralisations for verbs, and a compound identification module). These corrections are currently only applied for words that are not contained in a Dutch lexicon or in a name list, and at most one variant is considered.

We are still developing our implementation of TARSP and the SASTA modules. Our current implementation does not yet include the corrections that we experimented with for the Schlichting Appendix. The intermediate results on the development data range for *Recall* from 73.5 to 93.9, for *Precision* between 51.0 and 77.1, and for *F1-score* between 60.2 and 84.7. If we currently run the system on the held-out test data, we obtain figures for *Recall* between 75.8 and 76.4, for *Precision* between 61.5 and 63.3, and for *F1-score* between 67.9 and 69.3 The detailed actual figures for these and additional data will be reported in the presentation and the final paper.

10 Concluding Remarks and Future Work

We have presented *Sasta*, an application to analyse transcripts of spontaneous language. Though the *Sasta* application applies only to Dutch, the techniques described here can be applied to any language provided there is a parser for that language and a query system for querying the syntactic structures resulting from the parser. We observe that SASTA scores pretty well on the grammatical analysis of transcripts of spontaneous language sessions. We also found that corrections of deviant language not only improves the deviant parts but also the overall analysis. SASTA also often finds more examples for a grammatical phenomenon than human annotators (who often overlook instances), but the human annotators remain superior in precision. Whether the quality of the grammatical analysis is good enough to make the whole process more efficient remains to be seen. With the VKL we will carry out experiments (starting in September 2020) in which *Sasta* will actually be used in the clinical setting so that we can assess this and optimally integrate *Sasta* into the normal workflow procedures of the hospitals and clinics. In addition, we have secured funding for a small successor project (SASTA+) in which we will investigate more advanced methods of detection and correction of deviations, including cases in which all words in the utterance are correct and cases where multiple variants should be considered. The automatic corrections developed here can also be used to improve existing CHILDES CHAT annotation files, and we will report on a side result of this work, a program called *CHAMP-NL* (CHAT iMProver for Dutch) to improve and enrich existing CHAT files.

References

[Augustinus et al.2012] Liesbeth Augustinus, Vincent Vandeghinste, and Frank Van Eynde. 2012. Example-based treebank querying. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

[Ball et al.2012] Martin J. Ball, David Crystal, and Paul Fletcher, editors. 2012. *Assessing Grammar: The Languages of LARSP*. Number 7 in Communication Disorders across Languages. Multilingual Matters, Bristol.

[Bishop1984] D. V. M. Bishop. 1984. Automated LARSP: Computer-assisted grammatical analysis. *British Journal of Disorders of Communication*, 19(1):78–87.

[Bol and Kuiken1989] Gerard Bol and Folkert Kuiken. 1989. *GRAMAT: Methode voor het diagnosticeren en kwalificeren van taalontwikkelingsstoornissen*. Berkhout, Nijmegen.

[Bouma et al.2001] Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

[Boxum et al.2013] Elsbeth Boxum, Fennetta van der Scheer, and Mariëlle Zwaga. 2013. *Analyse voor Spontane Taal bij Afasie. Standaard in samenwerking met de VKL.* VKL, October. `https://klinischelinguistiek.nl/uploads/201307asta4eversie.pdf`.

[Crystal et al.1976] D. Crystal, P. Fletcher, and M. Garman. 1976. *The grammatical analysis of language disability*. Edward Arnold, London.

[Long et al.1996 2000] S.H. Long, M.E. Fey, and R.W. Channell. 1996–2000. Computerized profiling, versions 9.0.3-9.2.7 (ms-dos) [computer program]. Software, Department of Communication Sciences, Case Western Reserve University, Cleveland, OH.

[MacWhinney2000] Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.

[Odijk et al.2017] Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 23, pages 281–297. Ubiquity, London, UK. DOI: `http://dx.doi.org/10.5334/bbi.23`. License: CC-BY 4.0.

[Odijk et al.2018] Jan Odijk, Martijn van der Klis, and Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 46–55, Prague, Czech Republic, January 23-24. `http://aclweb.org/anthology/W/W17/W17-7608.pdf`.

[Oostdijk et al.2002] N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In M. González Rodriguez and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.

[Pander Maat and Dekker2016] Henk Pander Maat and Nick Dekker. 2016. Tekstgenres analyseren op lexicale complexiteit met TScan. *Tijdschrift voor Taalbeheersing*, 38(3):263–304.

[Pander Maat2017] Henk Pander Maat. 2017. Zinslengte en zinscomplexiteit. *Tijdschrift voor Taalbeheersing*, 39(3):297–328.

[Parisse et al.2012] Christophe Parisse, Christelle Maillart, and Jodi Tommerdahl. 2012. F-larsp: A computerized tool for measuring morphosyntactic abilities in French. In Martin J. Ball, David Crystal, and Paul Fletcher, editors, *Assessing Grammar: The Languages of LARSP*, number 7 in Communication Disorders across Languages, chapter 13.

[Schlichting2005] Liesbeth Schlichting. 2005. *TARSP: Taal Analyse Remediëring en Screening Procedure. Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar*. Pearson, Amsterdam, 7th edition.

[Schlichting2017] Liesbeth Schlichting. 2017. *TARSP: Taalontwikkelingsschaal van Nederlandse kinderen van 1-4 jaar met aanvullende structuren tot 6 jaar*. Pearson, Amsterdam, 8th edition.

[van Ierland et al.2008] Margreet van Ierland, Jeannette Verbeek, and Leen van den Dungen. 2008. *Spontane Taal Analyse Protocol. Handleiding van het STAP-instrument*. UvA, Amsterdam.

[van Noord et al.2013] Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

[Verbeek et al.2007] Jeannette Verbeek, Leen van den Dungen, and Anne Baker. 2007. *Spontane Taal Analyse Protocol. Verantwoording van het STAP-instrument, ontwikkeld door Margreet van Ierland*. UvA.

A Neural Parsing Pipeline for Icelandic Using the Berkeley Neural Parser

Þórunn Arnardóttir
University of Iceland
Reykjavík, Iceland
tha86@hi.is

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

Abstract

We present a machine parsing pipeline for Icelandic which uses the Berkeley Neural Parser and includes every step necessary for parsing plain Icelandic text, delivering text annotated according to IcePaHC. The parser is fast and reports an 84.74 F1 score. We describe the training and evaluation of the new parsing model and the structure of the parsing pipeline. All scripts necessary for parsing plain text using the new parsing pipeline are provided in open access via the CLARIN repository and GitHub.

1 Introduction

A parsed corpus has many applications but the process of making such a corpus manually can be long and time-consuming, making automatic parsers an appealing option. A corpus made with automatic parsing can never be as accurate as a manually corrected corpus but can produce a much larger corpus in a fraction of the time. The neural parsing pipeline introduced in this paper is practical within both linguistics and language technology. The pipeline can be used for parsing large amounts of texts and the resulting corpus can be used for research in Icelandic syntax, analyzing syntactic movements or changes in syntax over time. Within language technology, a parsing pipeline can be used for parsing sentences to decipher their meanings, creating a parsed corpus for training various software, and more.

The Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al., 2011; Rögnvaldsson et al., 2012) is a manually corrected Icelandic treebank which consists of one million words. By training the Berkeley Neural Parser using texts from IcePaHC, an Icelandic parsing pipeline, IceNeuralParsingPipeline (Arnardóttir and Ingason, 2020a), is created which takes in plain text and delivers the text parsed according to the IcePaHC annotation scheme. The parsing pipeline builds upon a previous pipeline released through CLARIN, IceParsingPipeline (Jökulsdóttir et al., 2019), which includes the Berkeley Parser, an older version of the Berkeley Neural Parser which is not based on a neural network. The new parsing model is a faster and more accurate parser, with an 84.74 F1 score compared to the 75.27 F1 score of the older parsing model. Crucially, the parsing pipeline is open-source, licensed under the MIT license and available at the Icelandic CLARIN repository[1] and GitHub.[2]

The structure of the paper is as follows. Section 2 focuses on previous work on Icelandic parsers and describes IcePaHC while Section 3 describes the Berkeley Neural Parser. In section 4, we describe the training process and the resulting model's accuracy is reported in section 5. Section 6 describes the updated parsing pipeline and finally, we conclude in section 7.

2 Related work and the Icelandic Parsed Historical Corpus

Language technology resources for Icelandic have grown in number over the last decade and as more resources become available, the making of further resources becomes more feasible. Two Icelandic non-data-driven parsers exist, IceParser (Loftsson and Rögnvaldsson, 2007), which was developed before

[1] http://hdl.handle.net/20.500.12537/17

[2] https://github.com/antonkarl/iceParsingPipeline

IcePaHC was created and is a shallow parser based on finite-state transducers, and Greynir's parser, which uses hand-written context-free grammar (CFG; Þorsteinsson et al., 2019). After IcePaHC became available, the first full phrase structure parser was developed, IceParsald, which utilizes the Berkeley Parser (Ingason et al., 2014) and is trained on IcePaHC. This parser was then further developed for use in the parsing pipeline discussed above. The Berkeley Parser preceding the Berkeley Neural Parser uses a probabilistic context-free grammar (PCFG; Petrov et al., 2006), which has been replaced by the state-of-the-art neural networks. Using neural networks instead of PCFGs or CFGs has resulted in more accurate parsers (Alberti et al., 2015; Chen and Manning, 2014), calling for an updated parsing pipeline using the Berkeley Neural Parser.

The data used for training the Berkeley Neural Parser is the Icelandic Parsed Historical Corpus (IcePaHC). It is a one-million-word, diachronic treebank which includes texts from the 12th to 21st centuries, distributed evenly (Rögnvaldsson et al., 2012). The texts included in IcePaHC have been manually corrected according to the Penn Parsed Corpora of Historical English (PPCHE) annotation scheme which uses labeled bracketing in the same way as the Penn Treebank. The part-of-speech tagset used is based on the one particular to the PPCHE, with minor changes to adapt it to Icelandic grammar. The annotation scheme employed in IcePaHC splits sentences into matrix clauses and marks their phrases. These phrases and their tokens are marked according to the tagset, which consists of 15 tags and their postdashial features.[3]

3 The Berkeley Neural Parser

The Berkeley Neural Parser is a constituency parser, relying on a repeated neural attention mechanism (Kitaev and Klein, 2018). The neural attention mechanism clarifies how information is transferred between different locations in a sentence. Different locations in a sentence can serve each other, both based on their positions and their contents, and these two types of attention are utilized in the parser with good results. Training the parser is possible when a gold corpus, such as IcePaHC, is available.

15 trained parser models for 11 different languages are distributed by the authors of the Berkeley Neural Parser, none of them being Icelandic. When training a model, several parameters can be used and tweaked. Three different word representation models are available: BERT, ELMo and fastText, BERT outperforming them both when used with the parser (Kitaev et al., 2019). BERT creates deep, contextualized word representations, delivering vectors for each word in the training set (Devlin et al., 2019). A few different BERT models are available, for example a multilingual model trained on 104 languages including Icelandic, but no models exclusively trained on Icelandic text have been published as of yet.

4 Training the model

IcePaHC consists of one million words in 73,012 matrix clauses. 80% of these clauses, 58,308 clauses, are used for training the parsing model, 10% for the development set and 10% for the test set, 7,302 clauses each. Since IcePaHC consists of data from different centuries (dated 1150–2008), an even distribution in the different sets is guaranteed by dividing every tenth part of the corpus between the training, development and test set.

Since IcePaHC tags are manually corrected, they can be used in training the model so that the parser predicts Part-of-Speech tags along with phrase structure. The cased multilingual BERT model is also utilized, having produced a more accurate parser for nine different languages (Kitaev et al., 2019). The neural network used for training consists of four layers, its learning rate is 0.00005, its batch size is 32 and the batch size during evaluation is 16. All computations were performed on resources provided by the Icelandic High Performance Computing Centre at the University of Iceland.

5 Evaluation

The parsing model is evaluated using EVALB, included in the Berkeley Neural Parser software. The 7,302 sentences of the test set are parsed using the parsing model and its output compared to the gold test set, delivering precision, recall, F-measure, complete match and tagging accuracy.

[3] A description of the tagset can be found at `https://linguist.is/icelandic_treebank/Tagset`

When trained using the multilingual BERT model, the parser achieves an 84.74 F1 score on the IcePaHC test set with 94.05% tagging accuracy. Its recall is 84.43%, its precision 85.07% and complete match is 46.61%. Some experiments were done with training a model without using word embeddings and using a different combination of training, development and test set. The highest accuracy reached without using word embeddings was an 82.18 F1 score. Changing the combination of data in the training, development and test set was also unfavorable. Using the oldest 80% of the data in the training set, the youngest 10% of the data in the test set and the 10% there in between in the development set, the model's accuracy dropped to a 77.01 F1 score. Reversing the data so that the youngest 80% of the data is in the training set, the oldest 10% in the test set and the 10% of data there in between in the development set proved to be more beneficial, reaching an 82.57 F1 score.

The parsing model is not only accurate, but also fast, as it is able to parse 228 sentences per second when run on NVIDIA Tesla V100 GPU on a Linux operating system. The model parses over five times the amount of sentences in the same amount of time when compared to the Berkeley Parser. When both parsers are run on two Intel Xeon E5-2680v3 CPUs on a Linux operating system, the Berkeley Neural Parser parses 4.8 sentences per second while the Berkeley Parser parses 0.85 sentences.

6 The Icelandic Neural Parsing Pipeline (IceNeuralParsingPipeline)

The parsing pipeline described in this article is based on iceParsingPipeline (Jökulsdóttir et al., 2019), with three changes. Instead of using Detector Morse for punctuation splitting, a tokenizer for Icelandic text, Greynir's Tokenizer (`https://github.com/mideind/Tokenizer`) is used for shallow tokenization. The software tokenizes the text, returning each sentence with its tokens separated. Using the tokenizer replaces two steps in the older pipeline. As mentioned, the Berkeley Neural Parser model is then used instead of the Berkeley Parser model in the parsing step.

The pipeline is structured as follows. First, the plain input text is tokenized and divided into sentences using Greynir's Tokenizer. Each sentence is then split up into matrix clauses using a matrix clause splitter, i.e. a sentence is split if a coordinating conjunction occurs. This step is illustrated in Figure 1, wherein a sentence has been split into two matrix clauses, shown in separate lines. The sentence translates to: "They believe in heaven and eternal life, and I'd preferably like to take up their religion", the coordinating conjunction *og* "and" marking the beginning of the second matrix clause. The splitting of sentences into matrix clauses is done because sentences in IcePaHC are divided into matrix clauses and each sentence parsed separately. The text is then parsed using the trained model. After having been parsed, the text is postprocessed in two steps. The first one consists of restoring dashes and removing extra labels and brackets created by the parser and in the second one, the text's format is changed. Before this step, each matrix sentence is displayed in a single line but to make the sentences more legible, they are formatted to conform to the IcePaHC format, wherein subphrases of a sentence are shown in separate lines.

Þeir trúa og himnaríki og eylífulífi,
og þeirra trúarbrögð vildi eg helst taka

Figure 1: A sentence split into two matrix clauses.

Because of the parsing model's speed, the pipeline can be used for parsing large corpora. Two treebanks have been created by using the pipeline, a historical one containing 2,7 million words, NeuralMIcePaHC (Arnardóttir and Ingason, 2020b), and a 500-million-word one containing mostly contemporary data, IceConTree (Arnardóttir et al., 2020).

7 Conclusion

In this paper, we have described an open-source parsing pipeline for Icelandic which includes a fast parser, the Berkeley Neural Parser. The parsing model is trained using the Icelandic Parsed Historical Corpus along with a BERT model and reports an 84.74 F1 score. Parsed text can be used in both language technology and research in syntax, in making a treebank or determining the phrase structure of a single

sentence. In creating a parsing pipeline which can accept plain text and deliver its parsed counterpart, the parsing process is made accessible for those not specialized in computer science or linguistics and the parsing speed opens up the possibility of parsing large texts.

Acknowledgements

We would like to thank the anonymous reviewers for helpful comments as well as the High Performance Computing Centre at the University of Iceland for providing the resources to run our experiments.

References

C. Alberti, D. Weiss, G. Coppola, and S. Petrov. 2015. Improved transition-based parsing and tagging with neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1359, Lisbon, Portugal, September. Association for Computational Linguistics.

Þ. Arnardóttir and A. K. Ingason. 2020a. IceNeuralParsingPipeline. CLARIN-IS, Stofnun Árna Magnússonar.

Þ. Arnardóttir and A. K. Ingason. 2020b. NeuralMIcePaHC. CLARIN-IS, Stofnun Árna Magnússonar.

Þ. Arnardóttir, A. K. Ingason, S. Steingrímsson, S. Helgadóttir, E. Rögnvaldsson, S. Barkarson, and J. Guðnason. 2020. The Icelandic contemporary treebank (IceConTree). CLARIN-IS, Stofnun Árna Magnússonar.

D. Chen and C. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

A. K. Ingason, H. Loftsson, E. Rögnvaldsson, E. F. Sigurðsson, and J. Wallenberg. 2014. Rapid deployment of phrase structure parsing for related languages: A case study of insular scandinavian. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

T. F. Jökulsdóttir, A. K. Ingason, and E. F. Sigurðsson. 2019. A parsing pipeline for Icelandic based on the IcePaHC corpus. In K. Simov and M. Eskevich., editors, *Proceedings of CLARIN Annual Conference 2019*, Leipzig, Germany.

N. Kitaev and D. Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July. Association for Computational Linguistics.

N. Kitaev, S. Cao, and D. Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy, July. Association for Computational Linguistics.

H. Loftsson and E. Rögnvaldsson. 2007. IceParser: An incremental finite-state parser for Icelandic. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, pages 128–135, Tartu, Estonia, May. University of Tartu, Estonia.

V. Þorsteinsson, H. Óladóttir, and H. Loftsson. 2019. A wide-coverage context-free grammar for Icelandic and an accompanying parsing system. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1397–1404, Varna, Bulgaria, September. INCOMA Ltd.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.

E. Rögnvaldsson, A. K. Ingason, E. F. Sigurðsson, and J. Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey, May.

J. Wallenberg, A. K. Ingason, E. F. Sigurðsson, and E. Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9. http://www.linguist.is/icelandic_treebankhttp://www.linguist.is/icelandic_treebank.

Annotating risk factor mentions in the COVID-19 Open Research Dataset

Maria Skeppstedt, Magnus Ahltorp, Gunnar Eriksson, Rickard Domeij
The Language Council of Sweden, the Institute for Language and Folklore, Sweden
firstname.lastname@isof.se

Abstract

We here describe the creation of manually annotated training data for the Kaggle task "*What do we know about COVID-19 risk factors?*". We applied our text mining tool on the "*COVID-19 Open Research Dataset*" to i) select data for manual annotation, ii) classify the data into initially established classification categories, and iii) analyse our data set in search for potential refinements of the annotation categories. The process resulted in a corpus consisting of 50,000 tokens, for which each token is annotated as to whether it is part of an expression that functions as a "risk factor trigger". Two types of risk factor triggers were annotated, those indicating that the text describes a risk factor, and those indicating that something could *not* be shown to be a risk factor.

1 Introduction

The COVID-19 Open Research Dataset (CORD-19) is a free resource with scholarly articles on viruses from the coronavirus family, and on related topics. The dataset currently contains around 40,000 full text articles. Associated with the dataset is the Kaggle COVID-19 Open Research Dataset Challenge (Allen Institute For AI, 2020), which consists of nine different tasks, all with the aim of extracting from the data what has been published regarding different COVID-19-related research questions.

We will here describe our process for creating manually annotated training data that can be used for one of these nine tasks, the task "*What do we know about COVID-19 risk factors?*". Our data annotation process was partly built on a topic modelling tool, *Topics2Themes*[1] (Skeppstedt et al., 2018), which is being maintained and further developed within the Språkbanken Sam and SWE-CLARIN infrastructures. The tool was used for i) *selecting* data for manual annotation, ii) *classifying* the data into initially established classification categories, and iii) *analysing* the text material in search for potential refinements of the annotation categories.

Similar qualitative text analyses have been conducted in previous research as a preparation for creating annotation categories in the medical domain (Mowery et al., 2012). However, we are not aware of any previous studies that have used a tool similar to Topics2Themes for this task. We will here showcase how such a tool can be applied in interdisciplinary research, i.e. to support annotation of medical texts.[2]

2 Method

For the "*What do we know about COVID-19 risk factors?*"-task, the participants are asked to find out what epidemiological studies report about potential risks factors for COVID-19. Our suggestion for how to approach this task is to train a model to recognise language expressions that are used for describing risk factors for diseases in general, i.e. expressions that we could call "*risk factor triggers*".

To construct training data for such a model, we first compiled a "Risk factor sub-corpus" by extracting 30,000 paragraphs from CORD-19 that contained a risk factor seed word. We used a list of 104 seed

[1] https://github.com/mariask2/topics2themes

[2] This work was supported by the Swedish Research Council (2017-00626).

words which we had compiled in the following manner: (i) We first constructed a list of the words (or sometimes bi- and tri-grams) that occurred in the call for the "*What do we know about COVID-19 risk factors?*"-task, and that we estimated would be good seed words for more general text about risk factors. These included, for instance, "risk factors", "factors", "co-infections", "co-morbidities", "high-risk", "pre-existing", and "susceptibility of". (ii) We, thereafter, expanded the list by adding synonyms from the Gavagai living lexicon (Sahlgren et al., 2016). (iii) Words in the list that occurred very often in the CORD-19 corpus were then removed from the seed list as unigrams and specified further with bi-grams. E.g., "factors" was removed and replaced by bi-grams such as "socio-economic factors" and "environmental factors". (iv) We finally read some of the paragraphs containing the seed words, in search for new words to add, and found words such as "more common among" and "more likely".

2.1 Selecting, classifying and analysing the data

As our time available for manual annotation was limited, we decided to focus our effort on text paragraphs with a content typical for the 30,000 paragraphs in the risk factor sub-corpus. With limited annotation resources, we are not likely to be able to catch outliers, or even moderately infrequent content, but we might be able to gather data for training a model that catches typical expressions. Finding topics that represent re-occurring content in a text collection, and creating automatic classes in this content, can be done in an unsupervised fashion, for instance by using topic modelling. For this task, we therefore used the previously mentioned topic modelling tool, Topics2Themes. We used the tool's ability to automatically find synonym clusters in the text with the help of word embeddings. We use embeddings pre-trained on biomedical text (gbrokos, 2018).

We configured Topics2Themes to use NMF topic modelling to try to find 20 topics in the data set. Since NMF is a randomised algorithm, which might produce slightly different results each time it is run, the algorithm was run 50 times, and only topics stable enough to occur in all re-runs were retained. This resulted in that nine stable topics were found. The topic modelling algorithm also returns which paragraphs and which terms that are most closely associated with the topics detected. For each of the nine topics, we read the 15 most closely associated paragraphs, and classified them according to if they contained a mention of a risk factor for a disease or not. For five of these nine automatically extracted topics, more than one of the top 15 paragraphs associated with the topic described a risk factor for a disease. For these topics, additional associated paragraphs were manually classified in the same fashion (around 70 more texts for each topic). A total of 419 paragraphs were manually classified, of which 150 were classified as describing risk factors for diseases.

In addition to classifying the texts as to whether they describe risk factors, the functionality in Topics2Themes for documenting re-occurring themes that are identified when manually analysing the texts was used. Typically, such manually identified themes represent re-occurring information on a more granular level than the automatically extracted topics. The aim of this classification was to analyse the texts from a language use point of view, in order to identify whether there were annotation categories in addition to "describing a disease risk factor" that would be relevant.

Finally, in order to provide annotations that could help a machine learning model to detect the language used for expressing risk factors, i.e. the "risk factor triggers", rather than to perform a text classification task, we extended the 150 paragraphs in which disease risk factors were described with additional token-level annotations. These annotations consisted of identifying which tokens that were "risk factor triggers", i.e. tokens used to indicate that something is a risk factor.

3 Results and Reflections

Figure 1 shows the four lists of the Topics2Themes tool. The second list shows the nine topics extracted by the tool, with the first topic, "Results of studies of co-morbidities and other risk factors", selected by the user (as indicated by the blue background). Blue lines connect this selected list element with terms associated with the topic (to the left), and paragraphs associated with the topic (to the right). A green label (Me) shows that the text has been annotated as describing a risk factor for a disease, whereas a yellow label (No) shows that the text has been annotated as not containing any information on risk factors.

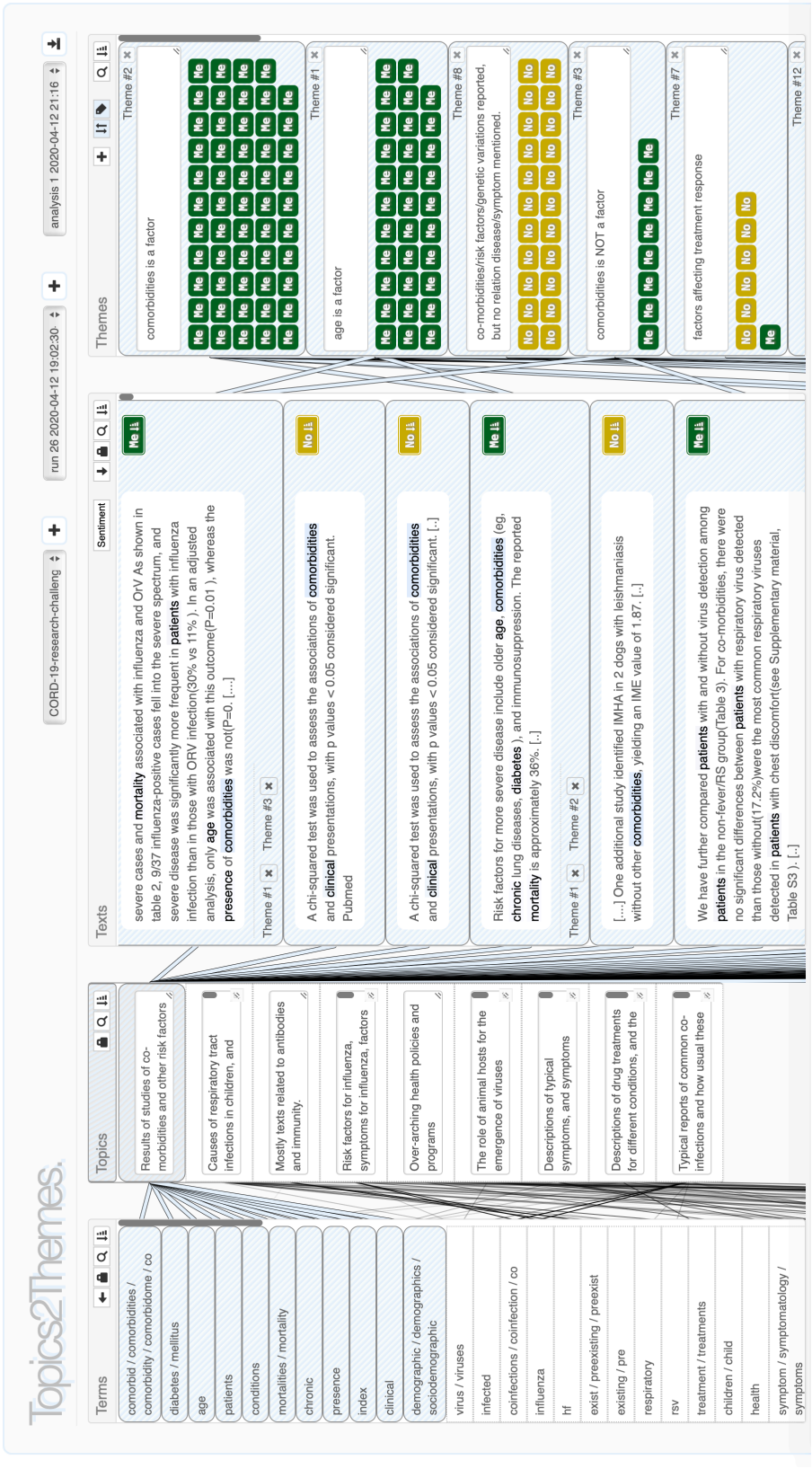


Figure 1: The second list from the right shows the nine automatically extracted *topics*. The user has selected the first topic "Results of studies of co-morbidities and other risk factors" by double-clicking on its list element. This has resulted in that the *terms* associated with this topic (to the left) – as well as the *texts* associated with this topic (to the right) – have been sorted as the top-ranked elements in their lists. The list of terms also shows the results of the automatic synonym clustering, based on word embeddings. Texts with a green label (Me) has been manually classified to contain mentions of risk factors, whereas texts with a yellow label (No) have been classified as not mentioning risk factors. The right-most list shows some of the *themes* that the user has identified when classifying the texts. The labels on the themes-elements show whether their associated texts have been classified as mentioning risk factors or not.

A total of 18 re-occurring themes were identified in the texts analysed, examples include "genetics (and family history) is a factor", "co-infection is a factor" and "age is a factor". Five additional examples are shown in the right-most panel in Figure 1. In 24 of the texts analysed, it is described that something could *not* be shown to be a risk factor. Five of the re-occurring themes identified described information of this type. Studies, in which it has *not* been possible to show that something is a risk factor, should be important to identify when mining for risk factors, since the information mined otherwise would be biased towards positive research results. We therefore decided to also include this information in the data to annotate, and consequently classified each such text with the green label that signifies that the text describes risk factors.

For the last annotation step, in which sequences of tokens were marked, we consequently created two annotation categories to differentiate between if the tokens functioned as (i) a risk factor trigger, or (ii) a trigger indicating that something could not be shown to be a risk factor. For instance, in the text: *"2019nCoV was of clustering onset, is more likely to infect older men with comorbidities [...]"*, the underlined text was annotated as a risk factor trigger. In contrast, the underlined text in *"The incubation periods did not significantly differ according to age, sex, or the presence of comorbidities [...]"* was annotated as a trigger describing that something could *not* be shown to be a risk factor.[3]

4 Future Plans

We have made our annotated data set, which consists of 50,000 tokens annotated for risk factor triggers, freely available at Kaggle.[4] We welcome anyone to use these annotations for training models. We would also appreciate efforts from others to annotate the same data set, in particular annotators with a medical background, as we have only been able to provide laymen annotations to the data.

The small size of the current data set, with only 150 positive samples, might not be sufficient for training a good model. We, however, plan to use these annotations as a seed set, and apply an active learning tool for further extending the annotated data set (Skeppstedt et al., 2016). At a later point in time, we also plan to create a data set that could be used for evaluating models trained on the type of data provided here. Such an evaluation set should be collected in a more controlled manner, e.g. through the annotation of all paragraphs in a few selected articles on the topic of COVID-19 risk factors.

References

Allen Institute For AI. 2020. COVID-19 open research dataset challenge (CORD-19). https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks.

gbrokos. 2018. Biomedical pre-trained word embeddings. https://github.com/RaRe-Technologies/gensim-data/issues/28.

Danielle L Mowery, Sumithra Velupillai, and Wendy W Chapman. 2012. Medical diagnosis lost in translation – analysis of uncertainty and negation expressions in English and Swedish clinical texts. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 56–64, Montréal, Canada, June. Association for Computational Linguistics.

Magnus Sahlgren, Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan, and Anders Holst. 2016. The Gavagai living lexicon. In *Proceedings of the Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Maria Skeppstedt, Carita Paradis, and Andreas Kerren. 2016. PAL, a tool for Pre-annotation and Active Learning. *JLCL*, 31(1):91–110.

Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018. Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.

[3]The IOB format was used for annotating the tokens, i.e. the entities in the examples were annotated as:
[... (is, **B-RISK**) (more, **I**) (likely, **I**) (to, **I**) ...**]** and **[**... (did, **B-NO**) (not, **I**) (significantly, **I**) (differ, **I**) (according, **I**) (to, **I**) ...**]**

[4]https://www.kaggle.com/mariaskeppstedt/manually-annotated-risk-factor-expressions

Contagious "corona" compounding in a CLARIN newspaper monitor corpus

Koenraad De Smedt
University of Bergen, Norway
desmedt@uib.no

Abstract

Newspaper corpora which are continuously kept up to date are useful for monitoring language changes in almost real time. The COVID-19 pandemic prompted a case study in the Norwegian Newspaper Corpus. This corpus was mined for productive compounds with the stems "corona" and the alternative "korona", tracing their frequencies and dates of first occurrence. The analysis not only traced the daily volume of such compounds, but also the sustained creation of many new compounds, and a change in their preferred spelling.

1 Introduction

Newspaper corpora are essentially time-stamped journalistic descriptions of daily events. Newspaper *monitor* corpora are continuously updated; thereby they not only provide a window into the course of current events, but they also provide an up-to-date data source for language use and language evolution in almost real time.

Such corpora are unfortunately scarce. In fact, the only monitor corpus that I could identify in the CLARIN resource family overview of newspaper corpora[1] is the Norwegian Newspaper Corpus (Andersen and Hofland, 2012) at the CLARINO Bergen Centre. This large (> 700 M words) and growing resource has proved useful in earlier studies of neologisms, loan words and other vocabulary expansion (e.g. Andersen and Graedler, 2020; De Smedt, 2012; Fjeld, 2012).

The COVID-19 pandemic, which spread rapidly from the spring of 2020, presented a good occasion for mining this corpus. In a relatively short time, the Norwegian vocabulary expanded drastically through the formation of thousands of new compounds. The most productive ones are based on the initial part *corona* or *korona*, such as *koronasituasjonen* ("the corona situation"), *corona-cruiset* ("the corona cruise") and *coronarasisme* ("corona racism"). What is quite unique about these compounds is that practically all those observed had probably never been used before 2020. In contrast to *virus* and other relevant terms, *corona/korona* is more specific and eye-catching, something that newspapers like and which may explain why its use seemed to have become contagious in the journalistic sphere.

Before 2020, the stem *corona* occurred in only a handful of relevant compounds such as *coronavirus, coronafamilien* ("the corona family"), and *corona-vaksiner* ("corona vaccines").[2] From January 2020, new compounds with *corona/korona* were increasingly observed, e.g. *koronatelefon* ("corona telephone"), *korona-frykt* ("corona fear") and *corona-situasjonen* ("the corona situation"). Productivity through the creation of new compounds, usually written as one word, sometimes with a hyphen, is generally high in Norwegian. This situation provided an exceptional opportunity to study the rate of productivity of a single word stem in detail.

The hypothesis of the present study was that there would not only be an evolution in the tempo of vocabulary expansion, but also that the corpus would show an evolution in the ratio between types and tokens. A secondary objective was to trace spelling change in terms of changing proportions of the two variant spellings. The hypothesis was that the normalization by the Language Council at the end of January 2020 would have an effect on journalists, but the extent of this effect had to be quantified.

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

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

2 Data

The Norwegian Newspaper Corpus (Andersen and Hofland, 2012) was the data source for the present study. It is updated every night by harvesting articles from ten major Norwegian online newspapers.[3] This corpus is accessible via an interface based on the IMS Corpus Workbench (CWB; Evert and Hardie, 2011).[4] The corpus is also accessible through the Corpuscle interface[5] (Meurer, 2012) at the CLARINO Bergen Center. The latter version has a better interface for query specification and download, but it is updated less regularly than the CWB version. Both versions were consulted, but the data from the Corpuscle version, which was up to date until May 26, 2020, are the basis for the present article.

The query `"[ck]orona.*" %c :: year = "2020"` was used in Corpuscle to retrieve all occurrences in 2020 of words starting with *corona* or *korona*, in uppercase or lowercase, from the Bokmål section of the corpus. Matching keywords were extracted, together with newspaper codes and dates, with May 26, 2020 as the last date. The base forms *corona/korona* and their inflected forms were removed, as well as obvious spelling errors and unrelated words such as *koronar* og *coronal*. The cleaned word list has 75240 tokens, which are all compounds, with or without hyphens.

This dataset was sorted, filtered and counted using a workflow of shell scripts and Awk programs. Lemmatization was done with the model *nb_core_news_md-2.3.0* in *Spacy*[6] and manually postprocessed to correct most lemmatizer errors. The results were aggregated and visualized with R scripts.[7]

3 Analysis

Before 2020, *coronavirus* and other rare compounds with *corona-* were only written with initial *c-*, whereas *koronautbrudd* ("corona outbreak") with *k-* was only used in the sense of "solar flare." In January 2020, the spelling with *c-* was still very dominant, while *k-* came into use from February. An article by the Language Council of Norway that recommended the spelling with *k-*, at a time when only the spelling with *c-* was in use, apparently caused the change. The present study is probably the first quantitative assessment of the effects of that normalization. Figure 1 shows the spelling variation from 2020 per day as well as between newspapers. After a brief period of fluctuation between spellings, majority use of *k-* was observed after the middle of February, as shown in Figure 1a. Furthermore, Figure 1b shows the variation per newspaper, revealing that a couple of newspapers mostly used the spelling with *c*.

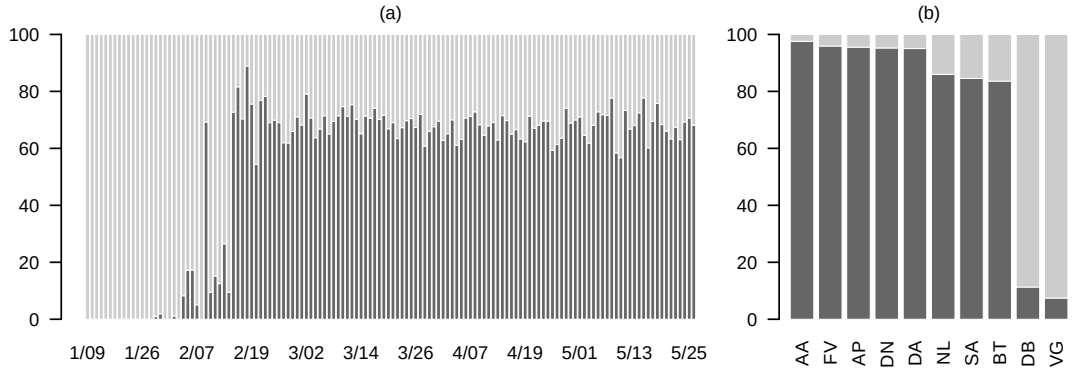


Figure 1. Percentages of tokens with *c-* (light) and *k-* (dark), (a) by date and (b) by newspaper

[3] Adresseavisen (AA, Trondheim), Aftenposten (AP, Oslo), Bergens Tidende (BT, Bergen), Dagsavisen (DA, Oslo), Dagbladet (DB, Oslo), Dagens Næringsliv (DN, Oslo), Fædrelandsvennen (FV, Kristiansand), Nordlys (NL, Tromsø), Stavanger Aftenblad (SA, Stavanger) and Verdens Gang (VG, Oslo).

[4] http://korpus.uib.no/avis/bokm.html

[5] http://clarino.uib.no/korpuskel

[6] https://github.com/explosion/spacy-models/releases//tag/nb_core_news_md-2.3.0, details at https://spacy.io/models/nb#nb_core_news_md.

[7] A frequency list and a list of new items by date of first occurrence can be accessed at https://github.com/clarino/corona.

The number of tokens per day is shown in Figure 2a. The earliest occurrences of relevant compounds in the Norwegian Newspaper Corpus in 2020 were *coronavirus* (indef. sg.) and *coronaviruset* (def. sg.), on January 20, 2020. Since that date, relevant compounds can be observed for every day in the studied period, which thus spans 139 days. The use of these and other compounds remained modest for over a month, but when the virus was detected in Norway, on February 26, a marked increase can be seen. The maximum frequency was 1752 tokens on one day.

The token counts are generally somewhat lower in the weekends when the volume of articles is lower. In that respect it might be useful to count normalized frequencies on the basis of the volume of harvested words per day, but unfortunately the daily volumes are not provided by the corpus interface. Token counts in themselves are however not the focus of the present investigation.

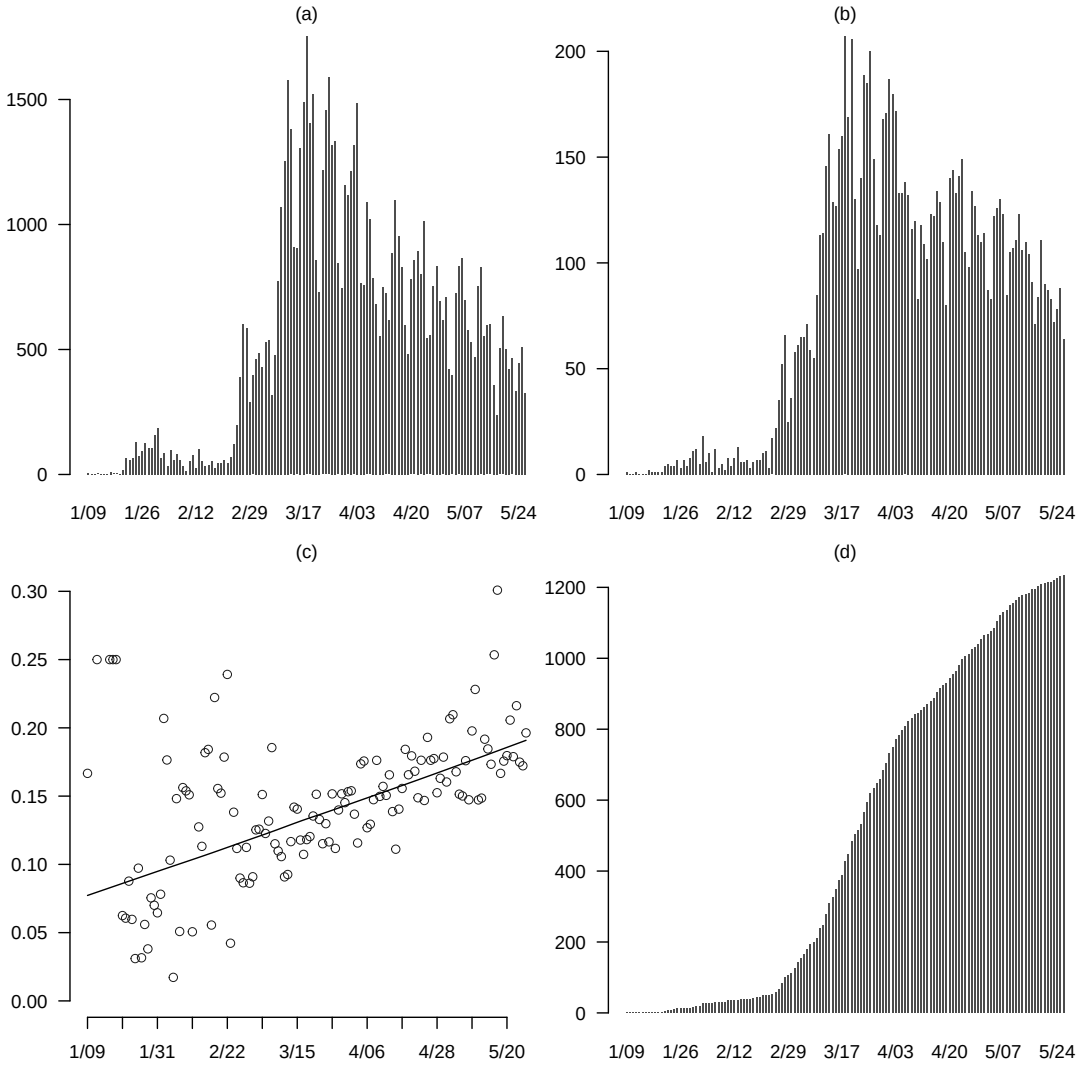


Figure 2. (a) Tokens, (b) types, (c) variation (ratio types/tokens) and (d) cumulative new types

While token volumes indicate how much is written about a topic in general, the breadth of the discussion in terms of subtopics may be revealed by looking at the number of distinct types. The 75240 tokens are distributed over 1235 lemma types (reduced from 1701 distinct word forms), disregarding the distinction between initial *c*- and *k*-, and disregarding the use of hyphens. A frequency list was made of all the types, showing a zipfian distribution. Compounds containing *virus* make up close to half of the total number

of tokens. Lemmatization produced some erroneous results and was not entirely consistent, e.g., some deverbal adjectives were reduced to the verb lemma, whereas others were not, but this did not seem to substantially affect the overall analysis.

The type count per day, as shown in Figure 2b, at first sight seems to roughly follow the increase in the token count. However, as Figure 2c shows with a trend line (fitted by local polynomial regression), the ratio of types to tokens per day was not constant, but was in fact increasing. This may be an indication of the fact that the variation in subtopics continued to grow markedly.

Another measurement is the number of *new* words per day, i.e. words which have not been recorded on earlier dates during this period (and not even before 2020, for almost all the words). A list was made of all final compound parts, with their first date of occurrence and the first newspaper in which they were observed. Counts of these new types were made per day. As Figure 2d shows, there is a steady increase in the vocabulary. In January and February the number of new words increased very slowly, but a sharp acceleration can be observed around February 26, when the virus had reached Norway. From that time on, there has been a steep and steady production of new compounds, with some flattening from April.

As expected, most compounds were nouns, e.g. *koronapsyken* ("the corona psyche"), some were verbs, e.g. *koronastenge* ("close down due to corona"), some were adjectives, including *koronafast* ("stuck due to corona", modeled after *værfast* "stuck due to bad weather" and *askefast* "stuck due to the ash cloud"), and some were adjectivally used participles, e.g. *corona-stanset* ("stopped by corona").

4 Discussion

This paper presents a use case demonstrating the potential of a newspaper monitor corpus, i.e. a corpus which is continuously updated with fresh newspaper articles, for the purpose of tracing and analyzing changes to the language in almost real time. In particular, this paper reports on the extraction and analysis of data from the Norwegian Newspaper Corpus. The course of new compound creation with *corona/korona* was tracked and analyzed for a period of 139 days. Due to limits on the size of this abstract, only a limited analysis is shown here. It was found that the majority spelling had changed in the course of about one month. Perhaps the most interesting result is that the variation of compounds in use, measured as the ratio of types vs. tokens, increased steadily. This is an indicator that a steadily widening range of situations and events was being reported. Finally, it was found that the creation of new compounds was a continuous process that accelerated from February 26.

The present findings show similarities and differences with a previous study on compounds with *aske* ("ash") following the volcanic eruption in Iceland in 2010 (De Smedt, 2012). That study found a sharp increase in variation but it was less broad in scope and it flattened out after half a month. In comparison, the presently reported increase in variation did not rise as quickly, but accelerated after about a month. Also, the creation of new compounds did not quickly decrease, but was sustained until the end of the studied period. Such characteristics may provide clues as to how fast, how broadly and how long different events affect our society, whether an event was initially underestimated, etc. This information may also benefit the automatic detection of significant "bursts" of words in information streams (Kleinberg, 2002, e.g.). Nevertheless, both in the earlier study and the current one, creative compounding seems to be contagious among journalists, who appear to outdo each other with ever more creative neology through compounding which is not necessarily transparent. Indeed, before 2020 it would have been difficult to interpret, for instance, *korona-telt* ("corona tent"), *koronautsettelsene* ("corona postponements") and *coronalov* ("the corona law"). The list of new compounds with their dates of first occurrence could be used for further semantic studies on the time line.

There are other systems for news tracking or for tracing new words. Among news trackers, the European Media Monitor[8] has a long-standing reputation. However, its interface is oriented towards topic tracking and alerts rather than detecting neologisms. The use of the Norwegian newspaper archive Atekst Retriever[9], based on daily harvesting from even more media sources than the Norwegian Newspaper Corpus, was briefly considered. However, Retriever is less suitable for linguistic research as it returns many

[8] https://emm.newsbrief.eu

[9] https://web.retriever-info.com/services/archive, accessed March 17, 2020

webpages with formatted text rather than a concordance or a table allowing easier collection of keywords together with relevant annotation (i.e. at least date and source). Furthermore, the earliest mention of a relevant compound with *corona/korona* in Atekst Retriever was December 28, 2019, which turned out to be a mistake in dating: in reality it was an article from February 28, 2020. It must be added, however, that also the Norwegian Newspaper Archive had some issues with dates, but these were either recoverable formatting errors or misplacements of lines which each still had correct dates. More generally, such issues show that although correct dating is paramount in studies like these, there may be errors that risk going under the radar; for a related discussion of hidden dangers in digitized text, see Nunberg (2009).

For the detection of neologisms, several systems are useful in their own right, such as The Word Spy[10] for English and Die Wortwarte[11] for German, the latter developed in the context of CLARIN-D and using monitor corpora. However, they seem to offer neither regular expression search, nor output as one complete list of observations with sources and dates. For those reasons, they are less suitable for the kind of data aggregation and analysis presented here. In fact, the Norwegian Newspaper Corpus also features a separate automated system which identifies new words every day. However, the goal of the present study is not so much to spot new words, but rather to trace both the creation and the protracted use of compounds based on a specific stem through a given period.

A study of *corona* compounds across languages might be interesting. However, despite the advantages of the availability of many newspaper corpora through CLARIN, the above-mentioned CLARIN resource family overview of newspaper corpora shows a lack of up-to-date monitor corpora. Almost all of the newspaper corpora in the CLARIN list consist of fairly dated materials and their different periods do not always overlap. Furthermore, the corpora are not easily interoperable; they do not share the same annotation and formatting, and they are not searchable through the same interface. Perhaps efforts should be made to implement Federated Content Search on this CLARIN resource family, and to promote the compilation of more monitor corpora that allow the study of vocabulary linked to current events.

Acknowledgements

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

References

Andersen, G. and Graedler, A.-L. 2020. Morphological Borrowing from English to Norwegian: The Enigmatic Non-Possessive -s. *Nordic Journal of Linguistics* 43(1):3–31.

Andersen, G. and Hofland, K. 2012. Building a Large Corpus Based on Newspapers from the Web. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1–28.

De Smedt, K. 2012. Ash Compound Frenzy: A Case Study in the Norwegian Newspaper Corpus. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins Publishing Company, Amsterdam/Philadelphia, 241–255.

Evert, S. and Hardie, A. 2011. Twenty-First Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics 2011 Conference*. Birmingham, UK.

Fjeld, R. V. 2012. Lexical Neography in Modern Norwegian. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins Publishing Company, Amsterdam/Philadelphia, 221–240.

Kleinberg, J. 2002. Bursty and Hierarchical Structure in Streams. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101.

Meurer, P. 2012. Corpuscle – a New Corpus Management Platform for Annotated Corpora. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Ed. by G. Andersen. Studies in Corpus Linguistics 49. John Benjamins Publishing Company, Amsterdam/Philadelphia, 31–49.

Nunberg, G. Aug. 31, 2009. Google's Book Search: A Disaster for Scholars. *The Chronicle of Higher Education*.

[10] https://www.wordspy.com

[11] https://wortwarte.de

Trawling the Gulf of Bothnia of News: A Big Data Analysis of the Emergence of Terrorism in Swedish and Finnish Newspapers, 1780–1926

Mats Fridlund • Daniel Brodén
Centre for Digital Humanities
Dept. of Literature, History of Ideas, and Religion
University of Gothenburg, Sweden
mats.fridlund@gu.se
daniel.broden@gu.se

Leif-Jöran Olsson • Lars Borin
Språkbanken Text
Dept. of Swedish
University of Gothenburg, Sweden
leif-joran.olsson@gu.se
lars.borin@gu.se

Abstract

This study combines history domain knowledge and language technology expertise to evaluate and expand on research claims regarding the historical meanings associated with terrorism in Swedish and Finnish contexts. Using a cross-border comparative approach and large newspaper corpora made available by the CLARIN research infrastructure, we explore overlapping national discourses on terrorism, the concept's historical diversity and its relations to different national contexts. We are particularly interested in testing the hypothesis that substate terrorism's modern meaning was not yet established in the 19th century but primarily restricted to Russian terrorism. We conclude that our comparative study finds both uniquely national and shared meanings of terrorism and that our study strengthen the hypothesis. In extension, the study also serves as an exploration of the potentials of cross-disciplinary evaluative studies based on extensive corpora and of cross-border comparative approaches to Swedish and Finnish newspaper corpora.

1 Combinining expertise in Language Technology and History

Partly thanks to the ongoing mass digitization of historical texts, historical research is currently entering an age of big data. Large-scale digital initiatives (LSDIs) to digitize historical texts provide novel ways for historians to analyze textual big data collections and address large-scale research questions. The development of language technology (LT) infrastructures allows historians and other scholars to trawl through massive amounts of text to quantitatively explore historical phenomena and to track conceptual changes over time. However, the use of LSDIs to address humanities research questions has been criticized for lacking in linguistic sophistication (see Zimmer 2013 and Tahmasebi et al. 2015) and in awareness of validity and representativity of the data sets. Tahmasebi et al. (2019) argue that many digital humanities projects are typically conducted with either strong data science or humanities bias, and are thus lacking in either appropriate humanistic domain knowledge to evaluate whether the results are pertinent or the technical methods suited to take advantage of the potential of the big data.

To get around such problems this study is part of an initiative of the Swedish CLARIN node, SweClarin, to foster interdisciplinary collaboration between researchers in humanities and LT, using LT-based e-science tools on big data textual infrastructures (see Viklund and Borin 2016 and Karsvall and Borin 2018). The paper builds on our previous study (Fridlund et al., 2019) that explored the usefulness to historical research of LT-based big-data methods to conduct large-scale corpus studies. The study combined domain expertise in history of terrorism with LT expertise in using text mining of an online newspaper archive to analyze the emergence of terrorism in a Swedish newspaper context 1780–1926. Through this we were able to evaluate and expand on prior historical research claims regarding terrorism's associated meanings and contexts in Sweden.

This paper extends the previous study through a comparative analysis of a similar Swedish-language corpus of newspapers published in Finland. For many centuries Finland was a part of Sweden. Swedish is still one of the two official languages of Finland with about 5% having it as mother tongue (50% bilingual) and the very first newspaper published in Finland was in Swedish. Following a war in 1809,

Finland was incorporated in the Russian Empire as a Grand Duchy, bringing it closer to the Russian political culture and its revolutionary context. In pursuit of overlapping discourses on terrorism and the historical diversity of the concept, our study will metaphorically trawl the 'Gulf of Bothnia' of digital historical newspapers. This gulf is the north arm of the Baltic Sea, consisting of Swedish and Finnish territorial waters and a shared body of water in between. Similarly, the body of texts we text-mine consists of specific domestic Swedish and Finnish news as well as a shared amount of transnational news appearing in both Swedish and Finnish newspapers. By identifying both overarching and more particular Swedish and Finnish meanings of terrorism, we will explore the gradual emergence of the current dominant modern meaning of terrorism and also how its historical diversity in meanings is related to different national contexts. The study will also serve to explore the potentials of evaluative studies based on extensive text corpora and of cross-border comparative historical approaches to newspaper corpora.

2 Language Technology resources and historical research questions

This study uses the corpus search tool Korp on historical Swedish-language newspaper corpora accessible through two national CLARIN B-centers: the National Swedish Language Bank (*Nationella språkbanken*) and the Language Bank of Finland (*Kielipankki/Språkbanken*). Korp (Borin et al., 2012) is a sophisticated corpus search tool with modular design with an online interface that allows searches and queries based on automatic linguistic annotations with structured result presentations: a *contextual hit list* or *KWIC* (keyword in context); *statistical data* of keyword occurrences in sub-corpora allowing creation of *trend graphs* plotting relative frequencies over time for text words, lemmas (dictionary headwords), or other linguistic items; a so-called *word picture* presenting statistically prominent fillers of selected syntactic dependency relations of a keyword. It should be noted that the two Korp configurations (seKorp and fiKorp respectively) are different. For example, data-wise the fiKorp gives an order of magnitude higher frequencies for some of the terms (Figure 1) partly due to better OCR quality. Due to differences in the features implemented in the two Korp configurations, multi-word comparison is not directly possible. However, we can nevertheless see from the trend diagrams that the peaks for individual terms show interesting tendencies warranting further investigation.

Our analysis period is chosen as 1780–1926 for pragmatic as well as historical reasons. The Swedish Kubhist is a large corpus of historical newspapers from late 18th to early 20th century digitized by the Swedish National Library, currently containing about 1 billion words (Adesam et al., 2019). While Kubhist covers the years 1749–1926, the corpus is not complete until 1780 and ends in 1926 due to copyright restrictions. It is well known that the meaning of terrorism diversified in the late 1870s to include political violence by Russian non-state militants and later broadened further to incorporate political substate violence outside of Russia. Our research questions concern the character of and changes in meanings and contexts of terrorism during the period. The meanings of terrorism that we investigate is what actors and violent political practices that were described as terrorists and terrorism and what particular ideologies that were attributed to terrorists, while the contexts of terrorism focus on the various national, physical and social spaces perceived by the newspaper media as habitats of terrorism and terrorists.

The investigation contributes to historical research on terrorism by expanding on our prior study (Fridlund et al., 2019) on the use of terrorism in Swedish newspapers, which indicates that although terrorist tactics were used by 19th century anarchists and anti-imperial militants, the concept's meaning was initially more specific and restricted than today, and not used more widely outside of the Russian context until the 20th century. The Swedish and Finnish cases are especially interesting in at least two contrasting and comparative ways. Sweden is with one bombing and one shooting 1908–09 a typical country for the period with no or few instances of terrorism, while Finland was a Russian Grand Duchy 1809–1917 with close proximity to Russian terrorist environments and suffered its own domestic terrorist campaign 1904–07, one of the earliest examples of explicit substate terrorism outside of Russia. Thus, a comparison of the two nations' terrorism coverage should reveal relevant similarities and differences. Furthermore, by combining historical domain expertise with LT competence, our analysis is an attempt to more generally demonstrate the gains of cross-disciplinary evaluative studies based on extensive corpora and in particular to inspire further cross-border historical studies of newspaper corpora.

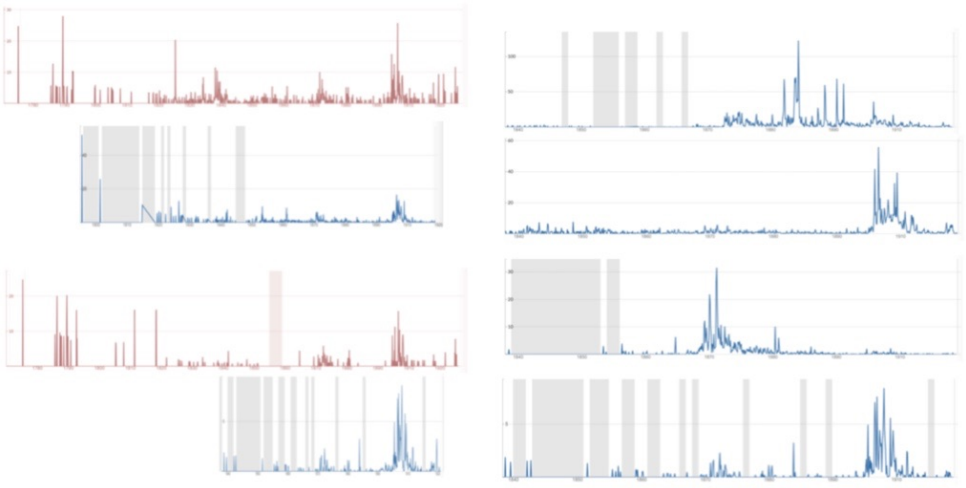


Figure 1: Trend graphs (with different scales) for two top left graphs showing *terrorist/terrorism* for seKorp (red) and fiKorp (blue) 1780–1926 and lower two *terrorist* for 1780–1920; and in right column (from top) *anarchist*, *revolutionary*, *nihilist*, and *terrorist* from fiKorp for the period 1849–1920.

3 Distant and close readings of terrorisms

To analyse the ideologies associated with terrorism in the Swedish and Finnish newspaper corpora the study investigates what historical meanings of terrorism were expressed in the national newspaper discourses and what political contexts the concept were primarily associated with. To include the wider context of political violence of which terrorism forms a part, we formulated our Korp queries combining the search terms *terrorist* (for 1780–1926 there were 259 and 1,364 hits in the Swedish and Finnish corpora respectively) and *terrorism* (570 and 2,361 hits), with closely associated terms used for actors who have been seen as making up a large part of the period's substate political militants such as *anarkist* 'anarchist' (3,028 and 20,837 hits), *nihilist* (1,660 and 3,113 hits), and *revolutionär* 'revolutionary' (noun: 1,285 and 9,618 hits). 'Nihilists' and 'nihilism' were – when not referring to modern philosophy – up until the 1890s often used as synonyms for Russian social revolutionaries and their ideologies.

Figure 1 shows trend graphs for central words related to terrorism practice during the period. The two top graphs in the left part are for *terrorist* and *terrorism* in seKorp (red) 1780–1926 and fiKorp (blue) 1805–1918, as there were no hits in the Finnish corpus before 1805 and after 1918. The two lower left graphs focus on *terrorist* for seKorp (red) 1780–1926 and fiKorp (blue) 1830–1918. In the right part fiKorp trend graphs for 1848–1920 describe the main militant actor groups. This period is chosen to take into account the European political upheavals of 1848 and the emergence from 1866 onwards of substate terrorism. The graphs' specific details are secondary to the comparisons of their relative profiles.

The following comparison between the terrorist-related contexts in Swedish and Finnish newspapers is an uneven comparison in that the fiKorp adaptation has not implemented the 'word picture' function for its Swedish-language material, as is the case for seKorp. To be able to compare the contexts we have used the results from the seKorp word pictures (as shown in Fridlund et al. 2019) and through a manual closer reading of the 3,725 concordances for *terrorist* and *terrorism* in the fiKorp KWIC view searched out these results for significant pre- and postmodifiers, such as nationalities, locations and gender.

The seKorp word pictures for *terrorist* and *terrorism* in Swedish newspapers produced several findings (reported in Fridlund et al. 2019) pointing to salient terrorist contexts of the period. For this study we will add a closer KWIC reading of the seKorp and some fiKorp hits and in some instances close reading of the actual newspaper articles the hits point to that can be read in an expanded KWIC context view or through a link in Korp to the original article in – for Swedish newspapers – the National Library of Sweden's digital newspaper archive *tidningar.kb.se* online or, if after 1903, physically on a restricted terminal at

the Humanities Library at University of Gothenburg, or for Finnish newspapers the National Library of Finland's digital collections (*Digi*) online at *digi.kansalliskirjasto.fi*.

The most significant findings relate to the *national contexts* – where terrorism came from – found through the national or ethnic attributes given to *terrorism* and *terrorist* in the newspapers. The most common in both seKorp and fiKorp are 'Russian' (*rysk*) terrorists and terrorism. Besides the Russian dominance it is difficult to generalize about how common various terrorist nationalities were. In addition to the 8 nationalities (Russian, Chinese, Finnish, French, German, Hungarian, Irish, and Polish) attributed in the seKorp word images, 14 unique nationalities were attributed in fiKorp (American, Armenian, Baltic, Bengali, Bulgarian, Czechian, Georgian, Indian, Italian, Latvian, Prussian, Romanian, Vatican, Wallachian). The only unique seKorp nationality was Chinese and several of the unique Finnish ones were part of the Russian empire. Several of these attributions referred to 19th century state terrorisms.

These national-ethnic attributions are significant in showing that terrorism was used both for (*state*) regime terrorism and (*non-state*) rebel terrorism. The 'Russian' occurrences point toward its well known rebel revolutionary terrorism campaigns of the 1880s and the early 1900s as well as state terrorism. The 'French' terrorism refers to the classical regime terrorism of the French revolution. Additionally, the findings contain several regime terrorisms in and by other national regimes 1848–67. The 'German' terrorism in seKorp was state terrorism by a Prussian army in the occupied Danish Duchy of Schleswig-Holstein in 1848. Similarly, the Hungarian and Polish terrorisms were connected to war and occupation where following the failed 1849 Hungarian revolution an occupying Austrian regime in Hungary in 1850 and a temporary rebel regime in Russian Poland in 1863 were accused of terrorism. This state terrorism in (contested) domestic or occupied foreign regions also account for several other regional terrorism attributions such as Armenian, Baltic, Czechian, Finnish, Grusinian, Indian, Latvian, Romanian, Svecoman, and Wallachian. State terrorism's prevalent existence is further supported by the country or regional locations of terrorism and terrorist in the seKorp word images where the only unique seKorp locations Croatia and Spain referred to state terrorism.

This anti-imperial terrorism is central in the emergence of a wider conceptualization of substate terrorism which, besides the Russian social revolutionary terrorisms of the 1880's and 1900's decades, is visible in three of the seKorp terrorist nationalities: Irish, Finnish and Chinese. The Irish terrorism appears 1882–89 and refers to both a local agrarian terrorism in the form of boycotts and murder – 'the agrarian murder' – of English settler farmers that spins off in an urban terrorist campaign. Those Irish urban terrorists are one of the earliest examples of non-Russians referred to as terrorists and make up the three classical terrorisms together with nihilists and anarchists. That said, this is a rare example before the 1900s of non-Russian substate terrorists. The Finnish nationality is especially interesting in showing how the definition of terrorism is expanding in the 20th century. Finnish terrorism occurs in two periods in the material, in 1905–06 and 1918. The first refers to both Finnish non-socialist and socialist terrorists, one of the first instances of the 'terrorist' label used outside of the immediate Russian context for describing a substate terrorism of the modern kind. The second Finnish terrorism emerges in the newly independent Finland in 1918 and is attributed to Finnish socialists that started the Finnish civil war. Finally, the 'Chinese terrorist' of a 1916 article was another non-Russian non-socialist revolutionary, a woman that had participated in the 1911 Chinese Xinhai Revolution and later was killed by the Chinese state.

Such anti-colonial and revolutionary terrorism is also visible in the Finnish newspapers that in 1907 reported about arrested 'Armenian terrorists' in Odessa and in 1909 that 'Indian terrorists' had renewed their secret activities. We find several 20th century references to anti-imperial terrorisms in the fiKorp material, especially connected to the Russian empire, besides Finnish and Armenian terrorists also 'Baltic' (1906), 'Latvian' (1907), 'Grusinian' (i.e. Georgian) (1912) terrorisms and substate and/or state terrorism occurring in empires and regions such as Turkey, Persia, Poland, and Macedonia.

The highly significant negative finding of the nationalities *not* attributed to terrorists (Fridlund et al., 2019) holds for fiKorp, although with an important modifier, i.e. that the spectacular terrorist deeds by anarchists in Germany, Italy, Spain, USA and the UK and by anti-colonial activists and separatists in Europe and Asia are not attributed to terrorists in Swedish or Finnish newspapers *during the 19th century*. There is only one anarchist attribution in an 1884 article published by two Finnish newspapers reporting

that Vienna had been put under a state of siege to make the population safe against 'the anarchists' terrorism' (*anarkisternas terrorism*). Besides that anarchist terrorism does not occur in the material although an article about Ireland in 1880 had warned about 'anarchic' (*anarkisk*) terrorism.

Our previous findings that terrorism was perceived as primarily an urban phenomenon were also strengthened by postmodifiers in the fiKorp KWIC views of European capitals (for instance London, Paris, Berlin, Stockholm) and other large cities (as well as some small ones). Also the previous noteworthy finding regarding the several 'female' (*kvinnlig*) synonyms attributed to terrorists got strengthened. It was also an occurring terrorist attribution in the Finnish press and referred to Russian socialist revolutionary terrorists in the 20th century – part of a Russian tradition of women terrorists going back to the 1870s – and which a close reading determined was also the case for the seKorp results.

4 Conclusions

Our study focuses on the opportunities for historians afforded by the Korp interface and by comparative and contrastive big data studies of national online newspaper archives. As expected and hoped for, our study could confirm what we knew from the historical record. But it also produced new findings about the changing diversity of the meaning of terrorism during the 19th and 20th centuries, especially demonstrated by state character attributions for the later period of the 19th century. Most importantly, the hypothesis that the modern meaning of substate terrorism was not yet widely established in the 19th century is strengthened by the results. Taken together a clearer and at times more complex image is emerging of the various salient state and substate terrorisms of the period, especially substate terrorisms so far relatively neglected by historians, such as the Macedonian, Armenian, Finnish and Indian terrorisms. In extension, our investigation points to the potentials for future more extensive cross-border historical studies drawing on Swedish and Finnish newspaper corpora.

Acknowledgements

The work presented here has been partly supported by an infrastructure grant to Språkbanken Text and Centre for Digital Humanities, University of Gothenburg, for contributing to building and operating a national e-infrastructure funded jointly by the Swedish Research Council (under contract no. 2017-00626) and the participating institutions.

References

Yvonne Adesam, Dana Dannélls, and Nina Tahmasebi. 2019. Exploring the quality of the digital historical newspaper archive KubHist. In *Proceedings of DHN 2019*, pages 9–17, Aachen. CEUR-ws.org.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Mats Fridlund, Leif-Jöran Olsson, Daniel Brodén, and Lars Borin. 2019. Trawling for terrorists: A big data analysis of conceptual meanings and contexts in Swedish newspapers, 1780–1926. In *Proceedings of Histo-Informatics 2019*, pages 30–39, Aachen. CEUR-ws.org.

Olof Karsvall and Lars Borin. 2018. SDHK meets NER: Linking place names with medieval charters and historical maps. In *Proceedings of DHN 2018*, pages 38–50, Aachen. CEUR-ws.org.

Nina Tahmasebi, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, Fredrik Johansson, Richard Johansson, Mikael Kågebäck, Olof Mogren, Pierre Nugues, and Thomas Risse. 2015. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2–4):169–187.

Nina Tahmasebi, Niklas Hagen, Daniel Brodén, and Mats Malm. 2019. A convergence of methodologies: Notes on data-intensive humanities research. In *Proceedings of DHN 2019*, pages 437–449, Aachen. CEUR-ws.org.

Jon Viklund and Lars Borin. 2016. How can big data help us study rhetorical history? In *Selected Papers from the CLARIN Annual Conference 2015*, pages 79–93, Linköping. LiUEP.

Ben Zimmer. 2013. When physicists do linguistics. Is English 'cooling'? A scientific paper gets the cold shoulder. *Boston Globe*, February 10.

Studying Emerging New Contexts for Museum Digitisations on Pinterest

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

Daniel Holmer
Department of Computer and Information Science
Linköping University, Sweden
danho775@student.liu.se

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

Arne Jönsson
Dept. of Computer and Information Science
Linköping University, Sweden
arne.jonsson@liu.se

Abstract

In a SweClarin cooperation project we apply topic modelling to the texts found with pins in Pinterest boards. The data in focus are digitisations from the Swedish History Museum and the underlying research question is how their historical objects are given new contextual meanings in the boards. We illustrate how topics can support interpretations by suggesting coherent themes for Viking Age Jewelry and localising it in different strands of popular culture.

1 Introduction

The content sharing platform Pinterest promotes itself as a visual discovery engine for ideas. It invites users to create themed collections called boards, either by linking images from other websites or by selecting among the images circulating on the platform, for an example see Figure 1. The platform serves an increasing number of users with images out of a growing bank of pins (at the beginning of 2020, 300 million users and 200 billion images). Once on Pinterest, images are set in motion by machine learning algorithms that present users with grids of images that change with each subsequent click.

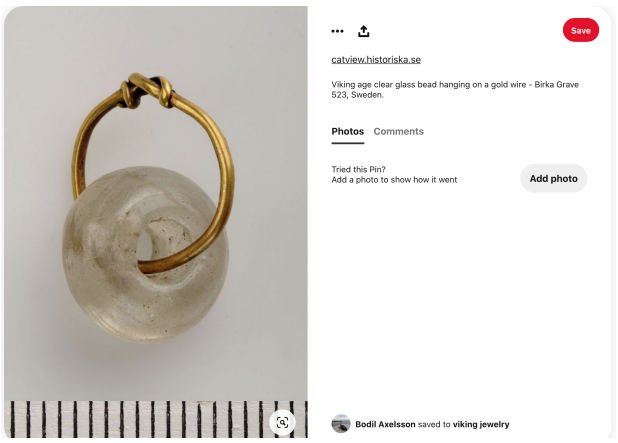


Figure 1: Typical pinterest entry used in this study

Digitisations, that is digital images of museum objects and art works, from all kinds and sizes of museums appear on the platform. Major international museums use the platform for dissemination. In addition, any computer savvy individual can source digitisations to Pinterest from blogs, websites or open collection management systems (Wilson, 2016). By offering a space for curation, the platform transfers agency from museum curators to the public in a space designed for consumption (Kidd, 2014).

The meaning of museum objects depends on available techniques and genres for interpretation, framing and contextualisation (Kirshenblatt-Gimblett, 1998). Historically, the contextualisation of museum

objects has been reserved to museum professionals. However, new museological practices and cultural policy directives have pointed to digital technology as a means to delegate contextualisation to audiences. Collection management systems initially created to keep track of objects and store information for professional purposes are now open to the public and collection data circulates online (Axelsson, 2019).

The aim of this paper is to investigate new contextual meanings of digitisations of Viking Age Jewelry from the Swedish History Museum on Pinterest. Digitisations from the Swedish History Museum that find their way to Pinterest are often recognizable as sourced from this particular museum. They may carry a source link back to the museum's database or to the museum's Flickr account. Moreover, they are depicted according to conventions for picturing the museum's objects, a view from above against a light greyish background, sometimes with a measure stick. In the museum's database, digitisations are embedded in knowledge developed within collection management and the discipline of archeology such as inventory number, find location (place, parish, region and country), estimated dating, substance, category, keyword and type. Importantly, only a selection of the museum's metadata is transferred to Pinterest. For each subsequent board the pin is saved to provide a new context (Hall and Zarro, 2012). When a user saves a digitisation it is sourced with url and a brief text description. The platform identifies images with a signature that is common for all subsequent repinnings to additional boards and links are created between all the boards that contain the sourced digitisation. Thus, there is an accumulative enrichment of the context for the image (Liu et al., 2017).

The first author made contacts with the Swe-Clarin K-centre at Linköping University to discuss methods for analysing the text descriptions. We agreed that topic modelling would be a suitable technique for the purposes of the project, at the same time furthering the K-centre's skills and the CLARIN goals of supporting scholars 'who want to engage in cutting edge data-driven research'[1].

2 Data

The starting point for the creation of data in this study was a qualitative immersive experience with the aim of understanding how digitisations from the Swedish History Museum circulate on Pinterest. The original query phrase that set the frames for selection was Viking Jewelry, a term that exists on Pinterest, but not in any museum database. Conditioned on the one hand by the researchers ability to recognize objects from the Swedish History Museum and relevant boards, and on the other hand the platform's recommendation system, the collection of data took place between March 2018 and October 2018. It stopped when the platform's recommendation service started to suggest boards already collected.

The approach taken can be described as data-intensive heritage ethnography with a mix of human centered interpretation and quantitative data-driven analysis (cf. Bonacchi and Krzyzanska (2019)). It takes turns between an interpretative approach and computational analysis. Data from 480 boards created in interaction with the platform was fetched by using Pinterest's developer API.

The dataset comprises a total of 329,999 entries. From this we filtered out duplicates, both description and picture duplicates, and entries with empty description fields, giving us a dataset of 107,165 unique entries. Data was tokenized with the NLTK tokenizer for English. The majority but not all descriptions are in English, we also identified Swedish, Russian, Norwegian, German and Dutch, but the English tokenizer was used for all languages. We filtered out words using the NLTK lists of stop words, with some domain specific additions such as *image*, *search*, and *show*, words that probably are auto generated by Pinterest. The texts were then lemmatized using the NLTK lemmatizer, and multi-word units such as *Thors hammer* were identified using bigrams and Gensim[2].

3 Topic modelling

The purpose of topic modelling is to reveal thematic patterns in a collection of documents. This enables extraction of knowledge from large collections of texts, that would otherwise be near impossible to do manually. Even though the main focus of each pin on Pinterest is its associated image and the platform's

[1] https://www.clarin.eu/content/vision-and-strategy

[2] `https://radimrehurek.com/gensim/`

recommendation system, users often provide a brief description of the image content. Semantic information is still vital for making relevant recommendations to users. It is easily extracted and is readable to both humans and machines as clues to what images represent (Zhang and Lapata, 2017). The pins are themselves assembled in different boards, where each board contains a certain category of pins, for example rings or pearls. In this study the aim was to discover the thematic patterns across a collection of boards which all related to viking jewellery one way or another. These thematic patterns are, after being extracted by a topic modelling method, represented as separate collections of keywords, or topics, which give an overview of the most prevalent words in each topic. Previous to the topic modeling, the 50 most common words were filtered out because they obscured the specificity of each topic. The extracted topics, and what they mean in the given context, can subsequently be interpreted by the analyst.

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

One of the main challenges when creating a competent LDA-model is to determine the number of topics, which has to be specified in advance of its creation. To aid in this process, a *coherence score* was used, which serves as a way to assess the semantic quality of a topic, and has shown to be largely correlated with human evaluations (Newman et al., 2010). The assumption is that a model with a high overall topic coherence score has topics that makes more sense to a human, than a model with a low overall topic coherence score. Gensim implements the framework proposed in Röder et al. (2015) to calculate coherence scores, and this was also used in our study. After calculating the coherence score on a wide range of numbers of topics, the number of topics that was found to give the highest overall score was 13, and was therefore chosen as the number of topics to base the model on.

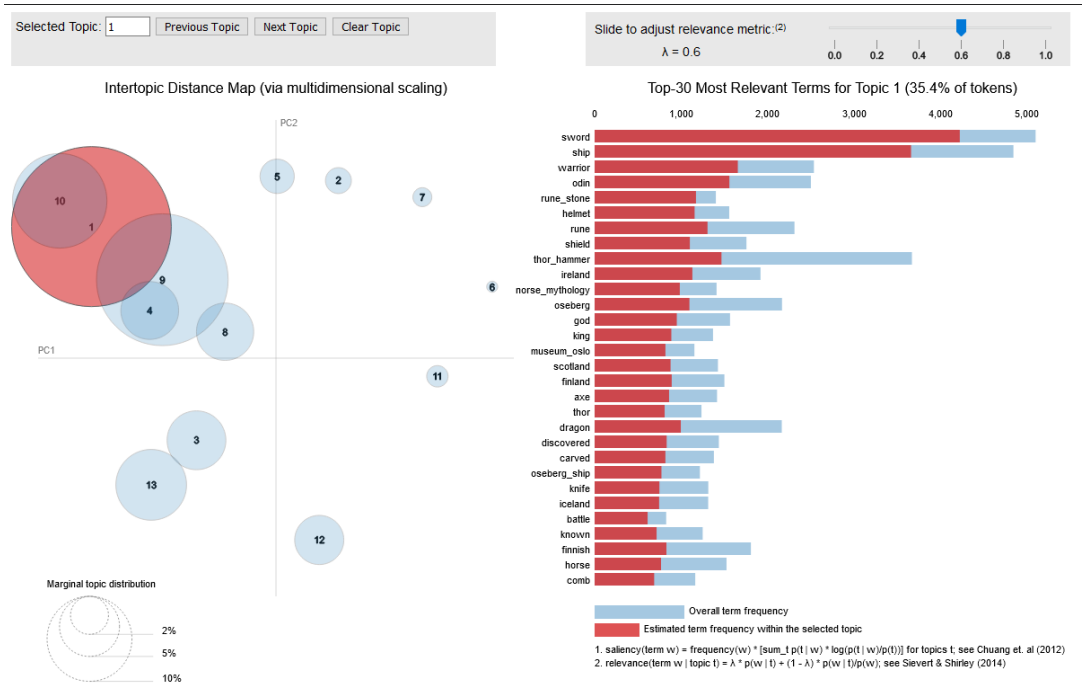


Figure 2: Visualisations of the topic models

The output of a LDA topic model using Gensim is a multinomial distribution over the topics and their

most prevalent words. To aid in the interpretation of the distributions we used a Python implementation[3] of LDAvis (Sievert and Shirley, 2014), which through a web interface visualize how prevalent each topic is, as well as how the contents of different topics relate to each other, see Figure 2. The topic clusters are projected as circles in a two-dimensional plane, where the relative sizes and distances between the circles represent prevalence and similarity of the topics. This means that a topic with a large circle is is seen more frequently in the entire collection of boards, and circles with a closer proximity share more features than circles that is projected further apart. In addition to the projection of the topics, there is also a bar chart of the most prevalent words in each topic. These are shown together with the frequency of the word in the entire corpus, allowing for a better understanding of the importance of the word in the current topic. LDAvis also introduces the term relevance, which is a way of ranking the words within the topics. By adjusting the relevance metric closer to 0 it is possible to filter out words that are globally frequent, and assigning higher weights to words that are unique to the topic, while a relevance metric closer to 1 uses the standard ranks of the LDA model. The assumption is that globally frequent words might be too common and not accurately reflect what differs between the different topics. However, being too strict with the filtering comes with a drawback; the unique words are innately rare, which often makes the topics hard to interpret. That is, a word that is relevant to a topic will be undervalued and ignored, if it appears in another topic simultaneously. This is of particular importance for this study, where the descriptions all revolve around vikings, and many relevant words are shared between topics. Sievert and Shirley (2014) therefore suggest a somewhat balanced relevance metric, about 0.6, which was also the value we found made the topics make the most sense when interpreting them.

4 Analysis

When proceeding with the analysis of the topics it is important to acknowledge that topic modelling differs from humanistic interpretation. It makes words into tokens, takes them out of context and structures the results in terms of standardized meanings (Binder, 2016). It follows that the word clusters the method produces cannot by themselves be taken as representations of new contexts for Viking Age Jewelry on Pinterest. Instead, the analysis has to proceed by returning to the ways in which topics are embedded in collection of images, i.e. boards. The fact that topic modelling is designed for identifying documents that are relevant for each topic make it possible to return to the images of the boards that displayed the highest frequencies of each topic. The visual inspection of these boards suggests that it is relevant to name the topics as below and discuss three main groups of topics. The first consists of four of the topics in the upper left corner of Figure 2. They are all tightly connected to how Vikings Age finds are explored in activities in which participants conjure up media images and museum objects to make costumes and props so as to re-live and reconstruct imagined pasts.

Topic group	Topic title	Most frequent keywords
I	Exploration, War and Mythology (1)	sword, ship, warrior, Oden, rune_stone, helmet
I	Northern European Tribes (4)	celtic, saxon, fibula, merovingian, celt, torque
I	Lajv, Fantasy and imagined worlds (8)	make, deviantart, armor, costume, dresss, cythian
I	Re-enactment and authenticity (9)	dress, apron_dress, embroidery, wool, tunic, coat
II	Rings (2)	Twisted, armring, finger_ring, gram_overall
II	Pearls (3)	Carnelian, roman, eye, bead_motiv, vikingtid_sted
II	Birka (5)	björkö_adelsö, thor_hammer, textile_fragment
II	Brooches (7)	oval_brooch, tortoise_brooch, bestillingsnr_lisens
II	Metal adornments (10)	thor_hammer, finger_ring, fibula, arm_ring, cast
II	Metal work (11)	tutorial, ear_cuff, knit, wire_wrapped, elf_ear
III	Tribes and trade (6)	columbia_river, antique_venetian, marble, african_trade
III	Shinies (12)	sterling_silver, earring, handmade, etsy, copper
III	Ancient jewelry styles (13)	roman, garnet, earring, byzantine, circa_century

Table 1: Topic titles after interpretation and the most frequent keywords for each topic

The topics in the second group may be regarded as subsets of the first because these topics consist of jewelry associated with Vikings (10), specific types of jewelry such as rings (2), pearls (3), and brooches

[3] `https://github.com/bmabey/pyLDAvis`

(7), finds from a particular excavation site in Sweden (5) or techniques for producing jewelry (11). Finally, there are three topics that have in common that the typical boards only contain a small number of items from the Swedish History Museum (Among them Figure 1). Here Viking finds are displayed together with jewelry from a multitude of periods and styles. As indicated in Table 1, many topics overlap, in terms of both keywords and images. They may only be distinguished as separate when taking account of all their tokens, visual inspection of boards and detailed in further analysis.

References

Bodil Axelsson. 2019. Breaking the frames: The creation of digital curatorial agency at swedish cultural historical museums (1ed.). In B. Brenna, H. D. Christiansen, and O. Hamran, editors, *Museums as Cultures of Copies.: The Crafting of Artefacts and Authenticity*, pages 239–252. London: Routledge.

Jeffrey M. Binder. 2016. Alien reading: Text mining, language standardization, and the humanities. In Matthew K. Gold and Lauren F. Klein, editors, *Debates in the Digital Humanities 2016*, pages 201–217. Minneapolis: University of Minnesota Press.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

C. Bonacchi and M. Krzyzanska. 2019. Digital heritage research re-theorised: Ontologies and epistemologies in a world of big data. *International Journal of Heritage Studies*, 25(12):1235–1247.

Catherine Hall and Michael Zarro. 2012. Social curation on the website pinterest. com. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–9.

Jenny Kidd. 2014. Museums in the new mediascape: Transmedia, participation. *Ethics*.

Barbara Kirshenblatt-Gimblett. 1998. *Destination Culture: Tourism, Museums, and Heritage*. Univ of California Press.

David C. Liu, Stephanie Rogers, Raymond Shiau, Dmitry Kislyuk, Zhigang Zhong Kevin C. Ma, Jenny Liu, and Yushi Jing. 2017. Related pins at pinterest: The evolution of a real-world recommender system. In *International World Wide Web Conference Committee (IW3C2), Perth, Australia*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, page 100–108, USA. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Fiona Wilson. 2016. Queens of collection and curation: Pinterest use in the society for creative anachronism. Master's thesis, School of Information Management, Wellington University, New Zealand.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Evaluation of a Two-OCR engine Method: First Results on Digitized Swedish Newspapers Spanning over nearly 200 Years

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

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

Ove Dirdal
Zissor
Oslo, Norge
ove@zissor.com

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

Abstract

In this paper we present a two-OCR engine method that was developed at Kungliga biblioteket (KB), the National Library of Sweden, for improving the correctness of the OCR for mass digitization of Swedish newspapers. We report the first quantitative evaluation results on a material spanning over nearly 200 years. In this first evaluation phase we experimented with word lists for different time periods. Although there was no significant overall improvement of the OCR results, the evaluation shows that some combinations of word lists are successful for certain periods and should therefore be explored further.

1 Introduction

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

Kungliga biblioteket (KB), the National Library of Sweden, is the central source for digitized Swedish newspapers, offering access to more than 25 million pages via the web service "Svenska dagstidningar".[1] The accuracy of the OCR system is therefore an important factor in order to maximize the access and usability of the digitized collections. To address this, KB, in collaboration with the Norwegian software company Zissor,[2] has implemented a novel OCR technique for combining two OCR engines: Abbyy and Tesseract. The two-OCR engine method has so far only been used as an internal testbed, awaiting proper evaluation and possible improvements. In 2019, KB embarked on an infrastructure project together with Språkbanken Text,[3] the Swedish Language bank at the University of Gothenburg, which is the coordinating Swedish CLARIN (Swe-CLARIN) node and CLARIN B Center, with the aim of evaluating and improving the results of the method (Dannélls et al., 2019).

In this paper we describe the two-OCR engine method and report the first quantitative evaluation performed against the ground truth material, spanning the years 1818-2011. We also report on an attempt to improve the OCR process by increasing the OCR's build-in vocabulary. By experimenting with different

[1] https://tidningar.kb.se/
[2] https://zissor.com/
[3] https://spraakbanken.gu.se/en

word lists for different time periods we take diachronic aspects into consideration and thereby, hope to adhere to linguistic change in the course of time (Springmann and Lüdeling, 2017). To our knowledge, this is the first use case study of evaluating and improving the OCR accuracy of Swedish newspaper texts for this period.

2 Two-OCR engine Method

The two-OCR engine method was developed in 2017 in cooperation between KB and Zissor. The method was designed to enable adjustment and control of some key parameters of the post-capture stage of the OCR process, including dictionaries and linguistic processing, to match typical features of the newspaper as a printed product, characteristics that in a historic perspective change over time, such as layout, typography, and language conventions. The working principle of the method is based on the evaluation and comparison between the results from two separate OCR engines: Abbyy,[4] version 11.1.16 and Tesseract,[5] version 4. The results from the two engines are analysed by applying rule-based voting principles whereby the best results on the word level are used to create a new compiled ALTO XML text file. There is no reliance on the internal confidence scores of these engines, the system is based on comparing the two ALTO XML files, word by word. The comparison and process of selecting between the two outputs of the OCR conversion is fully automated. It relies on a scoring model that was implemented based on the internal dictionaries of each OCR program. In addition to the ALTO XML files there are also statistics as to errors and verifications generated as Excel-files for each page. In this way, the consequences of modifying the parameters in the OCR process can be closely monitored. Figure 1 provides an overview of the OCR module architecture.

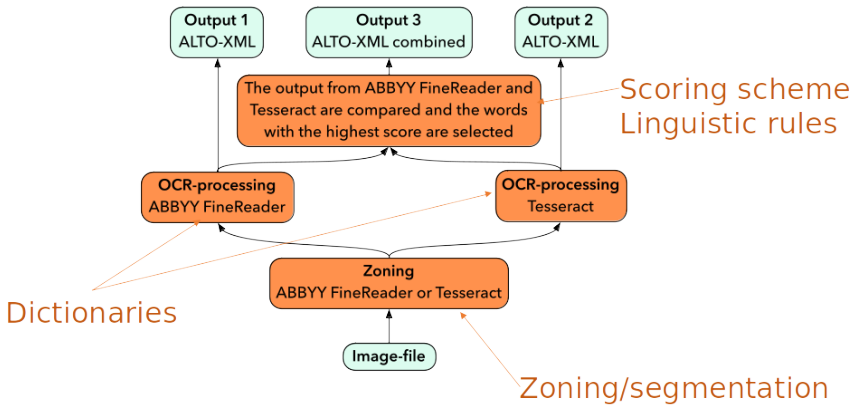


Figure 1: The design of the two-OCR engine method.

3 Reference Material and Processing

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

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

The selected reference material was sent to Grepect, a transcription company who specialises in double-keying.[6] Based on our inspections of the selected material we defined the guidelines for the

[4] http://finereader.abbyy.com/
[5] https://github.com/tesseract-ocr/
[6] http://www.grepect.de/

transcription which contained instructions for typeface, size and location changes. As the transcription process takes longer than expected we have begun to analyse the findings based on the presently available ground truth. So far we received 38 newspapers, covering the years 1818-2011 and amounting to 76 pages with 180024 words.[7]

To produce our baseline we run the material through the two-OCR engine system using Abbyy's and Tesseract's internal dictionaries exclusively. Next, we run the material in the two-OCR engine system four times, each run with a different word list (see Section 4). For each run the system delivers three results: one result for Abbyy, one for Tesseract and one calculated/verified Abbyy-Tesseract.

4 External Swedish Word Lists

We compiled four word lists from dictionaries and lexical resources from different time periods:

- Dalin, a full form lexicon for the 19th century, covering the morphology of late modern Swedish (Borin and Forsberg, 2011), containing 509,924 entries;
- Saldo, a full form modern lexicon (Borin et al., 2013), containing 1,704,718 entries;[8]
- Saol-hist,[9] a subset of the Swedish Academy historical lexicon, containing only base forms, amounting to 128,720 entries;
- Fem, a word list over name entities that was compiled from five lexical sources at KB, containing in total 311,481 entries.

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

Figure 2 shows the distribution of words in the word lists. As can be seen, the amount of words in newspapers from 1900-1999 dominate with a total of 136,077 words, of which 73,863 are found in Dalin and 91,603 in Saldo. The amount of words that are found in the pages from 1818-1899 is 29,942 of which 15,024 are found in Dalin and 15,875 in Saldo. When we inspected the list of words that were not found in any of the word lists we found that the majority of these words are numbers and non-alphanumerical characters.

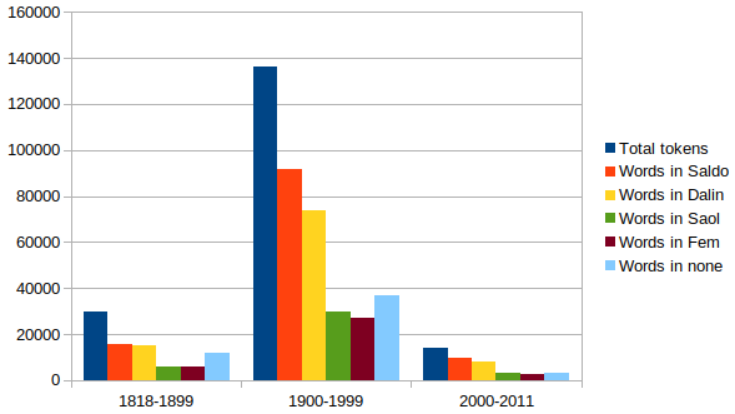


Figure 2: The distribution of words in the word lists for each time period complied from the ground truth.

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

[8]Both Dalin and Saldo are part of CLARIN lexical resources.

[9]`http://spraakdata.gu.se/saolhist/`

5 Evaluation Results

For evaluation we used the OCR frontier toolkit (Carrasco, 2014).[10] The method calculates the results of the OCR errors by measuring character accuracy rate (CAR) and word accuracy rate (WAR). We calculated the CAR and WAR results for each run against the prepared ground truth material of the 76 newspaper pages divided into three time periods. Table 1 shows the evaluation results of our runs without (baseline) and with external word lists.[11]

	1818-1899		1900-1999		2000-2011	
System	CAR (%)	WAR (%)	CAR (%)	WAR (%)	CAR (%)	WAR (%)
Abbyy baseline	81.51	61.7	94.74	93.26	94.05	93.42
Tesseract baseline	80.75	63.79	94.38	92.78	93.85	93.4
Abbyy-Tesseract baseline	81.64	65.23	94.78	93.52	94.22	94.01
Abbyy Dalin	80.98	**62.11**	94.63	93.14	94.05	**93.57**
Tesseract Dalin	**84.66**	**71.68**	93.83	92.6	93.57	92.65
Abbyy-Tesseract Dalin	81.13	64.73	94.65	93.24	94.17	93.94
Abbyy Saldo	81.08	**61.93**	94.6	93.03	93.99	**93.54**
Tesseract Saldo	**84.74**	**71.65**	93.82	92.6	93.56	92.67
Abbyy-Tesseract Saldo	81.22	64.45	94.6	93.09	94.06	93.74
Abbyy Saol-hist	80.91	**62.23**	94.67	93.22	94.03	**93.5**
Tesseract Saol-hist	**84.51**	**71.7**	93.82	92.6	93.58	92.67
Abbyy-Tesseract Saol-hist	81	65.13	94.68	93.33	94.18	93.92
Abbyy Fem	80.69	60.94	94.37	92.64	93.71	92.94
Tesseract Fem	**84.45**	**71.54**	93.82	92.60	93.57	92.66
Abbyy-Tesseract Fem	80.75	63.79	94.38	92.78	93.85	93.4

Table 1: Evaluation results of 76 pages run against the ground truth material with the two-OCR engine system, once without (baseline) and three times with external word lists, one run for each word list. The results marked in bold highlight successful runs that outperform the baseline.

As can be observed in Table 1, Abbyy shows a small improvement on word level with Dalin, Saldo and Saol-hist for 1818-1899. Interestingly, it also shows a small improvement on word level with these word lists for 2000-2011. There is an impressive improvement for Tesseract both on character and word levels with all word lists for the same period. It is striking that there is no improvement for Abbyy-Tesseract for the same period. Surprisingly, neither of the runs for 1900-1999 have improved over the baseline for Abbyy, Tesseract or Abbyy-Tesseract. This could be explained by the fact that the systems are eager to find a lexical match in the external word lists, and since the external lists get higher priority, wrong words are being replaced. Thus, there is a higher percentage of words that are replaced with incorrect ones for that particular context. Consequentially, CAR is also decreasing. Another explanation of the low performance is the high ratio of out-of-word vocabulary for this period as seen in Figure 2.

6 Related Work

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

Earlier work on historical English demonstrated the challenges with combining multiple OCR engines (Lund et al., 2011). They reported an improvement over the WAR using voting and dictionary features.

[10] `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.216.9106&rep=rep1&type=pdf`

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

Reul et al. (2018) applied a confidence voting scheme between OCR models that were trained on a single engine. Their evaluation on Latin books showed a relative improvement over the CAR.

OCR errors are often classified into two groups: non-word errors and real-word errors (Mei et al., 2016; Nguyen et al., 2019). The advantages of increasing the engine's vocabularies in order to correct real-word errors has been studied by several authors who have proven the usefulness of lexicons, among the successful strategies for improving the OCR accuracy (Kissos and Dershowitz, 2016; Schulz and Kuhn, 2017; Nguyen et al., 2018). Authors have shown that a lexicon-based approach is competitive if it is adapted to certain domains or time periods. Our assumption here is that curated word lists are suitable to experiment with for Swedish material that spans over hundred years since a great number of changes in orthography and morphology occurred during this time, in particular around 1906.

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

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

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

7 Conclusions and Future Work

We explored the effect of adding curated word lists to improve the two-OCR engine system when digitizing Swedish newspapers from nearly 200 years. We found that the addition of word lists in combination with the two-OCR engine system did not provide the expected significant improvement of the OCR result, even though some combinations proved more successful than others. There are however some unexpected variations in accuracy that have to be examined in detail in relation to the specific word lists and the given time period of the sample. The quality control of the two engine system verifies that the word lists are taken into consideration during the OCR process but their apparent unpredictable effect on the results have to be further analysed. The external word lists had noticeable effect on the internal confidence values of the OCR programs. The effect of these combined with the possibility to include the rate of correspondence between the results from the two-OCR engine (on the word level), could be used as a variable to indicate the quality of the results, something we will explore later in the project. Another important variable seems to be the scoring scheme which is governed by a set of rules for deciding on which of the systems processed the correct word. The impact of the rule set will be analysed once the consequences of the use of word lists is further examined.

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

Acknowledgements

The research presented here is funded by Riksbankens Jubileumsfond, the Swedish Research Council (grant agreement IN18-0940:1). It is also supported by Språkbanken Text and Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) Swedish CLARIN (grant agreement 821-2013-2003).

References

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

Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In *Language technology for cultural heritage*, pages 41–61. Springer:Berlin.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Rafael C. Carrasco. 2014. An open-source OCR evaluation tool. In *Proc. of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 179–184, NY, USA.

G. Chiron, A. Doucet, M. Coustaty, M. Visani, and J. Moreux. 2017. Impact of ocr errors on the use of digital libraries: Towards a better access to information. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4. IEEE.

Simon Clematide and Phillip Ströbel. 2018. Improving OCR quality of historical newspapers with handwritten text recognition models. In *Workshop DARIAH-CH*.

Dana Dannélls, Lars Björk, and Torsten Johansson. 2019. Evaluation and refinement of an enhanced ocr process for mass digitisation. In *Proceedings of Digital Humanities in the Nordic Countries*, pages 112–123. CEUR.

Senka Drobac, Pekka Kauppinen, and Krister Linden. 2019. Improving ocr of historical newspapers and journals published in finland. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 97–102. ACM.

I. Gregory, P. Atkinson, A. Hardie, A. Joulain-Jay, D. Kershaw, C. Porter, and C. Rupp. 2016. From Digital Resources to Historical Scholarship with the British Library 19th Century Newspaper Collection. *Journal of Siberian Federal University, Humanities and Social Sciences*, 9(4):994–1006.

I. Kissos and N. Dershowitz. 2016. OCR error correction using character correction and feature-based word classification. In *12th IAPR Workshop on Document Analysis Systems DAS*, pages 198–203. IEEE.

Mika Koistinen, Kimmo Kettunen, and Tuula Pääkakkönen. 2017. Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur and Antiqua Models and Image Preprocessing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA*. Association for Computational Linguistics.

W. B. Lund, D. D. Walker, and E. K. Ringger. 2011. Progressive Alignment and Discriminative Error Correction for Multiple OCR Engines. In *International Conference on Document Analysis and Recognition*, pages 764–768. IEEE.

Jie Mei, Aminul Islam, Yajing Wu, Abidalrahman Mohd, and Evangelos E Milios. 2016. Statistical learning for OCR text correction. *arXiv preprint*, abs/1611.06950.

Thi-Tuyet-Hai Nguyen, Mickaël Coustaty, Doucet Antoine, and Nhu-Van Nguyen. 2018. Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction. In *20th International Conference on Asia-Pacific Digital Libraries, ICADL*, November.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 2019. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL '19, page 29–38. IEEE Press.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickaël Coustaty, and Antoine Doucet. 2020. Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Virtual Event*, pages 333–336. IEEE.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning. *J. Lang. Technol. Comput. Linguistics*, 33(1):3–24.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored ocr post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2716–2726, Copenhagen, Denmark, September. Association for Computational Linguistics.

Uwe Springmann and Anke Lüdeling. 2017. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, 11(2).

Stimulating Knowledge Exchange via Trans-national Access – the ELEXIS Travel Grants as a Lexicographical Use Case

Sussi Olsen
University of Copenhagen, Denmark
saolsen@hum.ku.dk

Bolette S. Pedersen
University of Copenhagen, Denmark
bspedersen@hum.ku.dk

Tanja Wissik
Austrian Academy of Sciences, Austria
Tanja.Wissik@oeaw.ac.at

Anna Woldrich
Austrian Academy of Sciences, Austria
Anna.Woldrich@oeaw.ac.at

Simon Krek
Jozef Stefan Institute, Slovenia
simon.krek@ijs.si

Abstract

This paper describes the intermediate outcome of one of the initiatives of the ELEXIS project: Transnational Access. The initiative aims at facilitating interaction between lexicographers/researchers from the EU and associated countries and lexicographical communities throughout Europe by giving out travel grants. Several of the grant holders have visited CLARIN centres, have been acquainted with the CLARIN infrastructure and have used CLARIN tools. The paper reports on the scientific outcome of the visits that have taken place so far: the origin of the grant holders, their level of experience, the kind of research projects the grant holders work with and the outcomes of their visits. Every six months ELEXIS releases a call for grants. So far 23 visits have been granted in total; 13 of these visits have been concluded and the reports of the grant holders are publicly available at the ELEXIS website.

1 Background and Motivation

Even though lexicography has a long history of international research conferences, it has traditionally been a research area with limited knowledge exchange outside of each lexicographical institution, and in many cases lexicographic data has only been accessible to researchers from the institution who created the data and held the copyright. This tradition is partly related to the fact that practical lexicography has a strong commercial basis; lexicographical data used to be good business. But it also relates to the fact that enabling easy access to restricted data requires significant effort into facilitating and controlling this access - which again requires time and money not easily found in the budgets of lexicographic projects.

To this end, an important objective of the ELEXIS project is to stimulate knowledge exchange between lexicographical research facilities, infrastructures and resources throughout Europe, which can consequently mutually benefit from the vast experience and expertise that exist in the community. Inspired by other EU projects such as EHRI[1], RISIS[2], InGRID[3], and sobigdata[4], ELEXIS offers trans-national access activities in the form of visiting grants that enable researchers, research groups and lexicographers to work with lexicographical data which are not fully accessible online. Furthermore, grants offer access to professional on the spot expertise in order to ensure and optimise mutual knowledge exchange. Finally, grant recipients can gain knowledge and expertise by working

[1] https://ehri-project.eu/ehri-fellowship-call-2016-2018
[2] http://datasets.risis.eu/
[3] http://www.inclusivegrowth.eu/visiting-grants
[4] http://www.sobigdata.eu/access/transnational

with lexicographers and experts in NLP and artificial intelligence. The CLARIN infrastructure is one of the important infrastructures for these travel grant visits.

The trans-national access activities are expected to have a long-term impact specifically but not only for lesser-resourced languages, boost the network and infrastructure of the European lexicographic community, and facilitate future collaboration and knowledge exchange.

The trans-national activities represent a way of ELEXIS to enable access to restricted data, which has so far not been available outside of the hosting institutions, to researchers from other institutions and countries. As the results of research conducted in trans-national activities become available under open-access licenses, the international lexicographic community will become acquainted with previously inaccessible resources.

2 The Grants

The transnational activities consist of visiting grants of 1 to 3 weeks for researchers to experiment with and work on lexicographical data in a context of mutual knowledge exchange with the hosting institutions. Around five visiting grants are made available twice a year during the entire project period, amounting to 7 calls, i.e. 35-40 grants in total.

The following lexicographic institutions accept transnational visits during the ELEXIS project: Jozef Stefan Institute, Institute for Dutch Language, Austrian Academy of Sciences, Belgrade Center for Digital Humanities, Institute of Bulgarian Language Lyubomir Andreychin, Hungarian Academy of Sciences, K-Dictionaries, Det Danske Sprog- og Litteraturselskab, University of Copenhagen, Trier Center for Digital Humanities, Institute for Estonian Language, Real Academia Española.

Out of the 11 countries, where the hosting institutions are located, 8 countries participate in CLARIN and 5 out of the 11 hosting institutions are operating CLARIN B Centres.

2.1 The Calls

Researchers and lexicographers within the EU member states and associated countries are invited to apply for a visit of free access to and support from one of the lexicographical institutions.

The calls for applications include descriptions of the institutions and the lexicographical resources, tools, and expertise that are made available for the visitors. Researchers and lexicographers interested in visiting a particular host institution are encouraged to make motivated applications describing their background, the purpose of the visit etc.

2.2 Dissemination and Reporting

The calls are disseminated through ELEXIS website[5], mailing lists, newsletters (e.g. CLARIN Newsflash), social media and via leaflets at conferences. Particular effort was invested in disseminating the ELEXIS travel grants via social media. For the first call, not only the call was advertised but also each hosting institution was presented in a separate post. Furthermore, we accompany the grant holders from the announcement through their travel visits with social media posts as well as website portraits. In total, we have published 129 posts related to the first three calls.

Besides the website portraits[6] that are published before the research visits in form of a written interview, the final reports of the grant holders are published at the ELEXIS website and on social media.

3 Status after four Calls

After four calls, 13 visits have been completed, one of these 'as a virtual visit' and 10 visits have been postponed due to the uncertain situation with the current pandemic. The next call has been postponed till October for the same reason.

[5] https://elex.is/grants-for-research-visits/
[6] https://elex.is/category/grants-for-research-visits/

The winners of the first four calls come from institutions located in the countries shown in figure 1. The fact that the applications received originate from a wide range of countries proves that the transnational access (TNA) program has indeed reached out to the strong European lexicographic community as well as to communities that do not have an equivalent strong infrastructure.

After four calls, most of the infrastructures have had one or more visits. Some infrastructures are much more applied to than others. However, since each host has a fixed budget for approximately three visits, infrastructures that have reached their maximum number of visits and spent their budget, are left out of the list of hosting infrastructures for upcoming calls. In order to make sure that visits are somewhat evenly distributed among infrastructures, there have been cases where the TNA Committee, who selects the winners of each call, has given priority to applications addressing less popular infrastructures on condition that these applications were of sufficient quality.

In order to investigate their research experience, we divided the grant winners into expertise groups, as shown in figure 2. The grant holders represent a good mixture of different levels of experience. At their research visit most grant holders, experienced or not, prove to have limited or little previous experience in the specific fields of their proposed projects. Not surprisingly, most of the more experienced researchers applied with projects aiming to improve their individual skills in specific areas hitherto not part of their research practice.

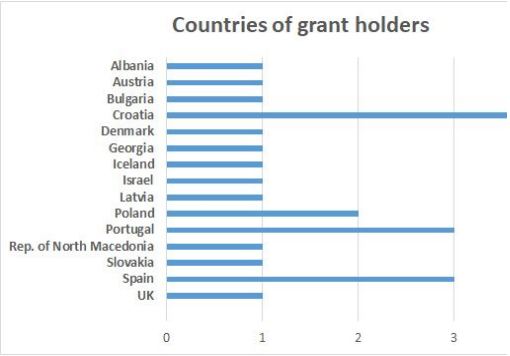


Fig. 1: Countries of grant holders

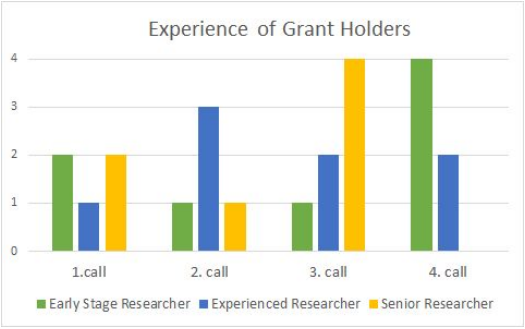


Fig. 2: Experience of grant holders

4 Scientific Outcome of the Visits

4.1 Research Visit Projects

The topics of the travel grant projects are quite diverse. Most of the projects focus on the compilation of dictionaries of different kinds and the primary objective of most visits are to be acquainted with the hosts' dictionary writing systems, corpus tools, the methodology behind the dictionary and corpus compilation, standards such as TEI and ISO, and to discuss their own project with experienced lexicographers and terminologists, see also Olsen & Pedersen 2020.

The projects span from retro-digitalization of older dictionaries, creation of dictionaries with terminological content, comparison of dictionary structures of different kinds, optimization of methods for automatic data extraction from corpora, the development of a data visualisation map and a study of business models of lexicography. Another topic concerns how to collect a well-balanced corpus for lemma selection. Many of the grant holders are very interested in gaining knowledge of the TEI guidelines for dictionaries, the ISO standards for lexicography and terminology as well as of dictionary encoding in XML, i.e. technical knowledge. In addition, a topic of current interest is the focus on the dictionary content in regard to ethical dilemmas in dictionary writing and potentially offensive content.

In their reports, all the grant holders emphasize the importance of communication and discussions with the experts of the hosting institutions. Several participated in workshops and other events and everybody reports of an instructive and rewarding visit.

4.2 Further Dissemination Related to the Research Visit projects

Besides the research visit and their published travel grant reports, the travel grant holders of the first call had the opportunity to present their projects during the poster session at the ELEXIS Observer Event in Vienna 2019. Furthermore, several of the research visit projects have led to scientific presentations, publications or were part of a Master thesis (cf. Woldrich and Wissik, 2019, 2020).

4.3 Benefits for the CLARIN Network

Some of the ELEXIS travel grant recipients come from countries that are not yet part of CLARIN, and are introduced to CLARIN during their visit, shown by the examples below.
The travel grant holder from Spain, who was visiting the Austrian Academy of Sciences in December 2019, was introduced to CLARIN AT and to all the benefits that CLARIN offers to researchers. Consequently a visit to the CLARIN K-Centre for Terminology Resources and Translation Corpora at the University of Vienna was organized. This was a perfect opportunity to exchange knowledge in the field of terminology with researchers involved and to learn more about the CLARIN K-Centre infrastructure.

A Croatian grant holder that visited the Society for Danish Language and Literature and University of Copenhagen was introduced to the CLARIN-DK infrastructure and CLARIN EU. With the help of CLARIN-DK staff, NLP tools from the CLARIN-DK toolbox, i.e. lemmatiser and pos-tagger, were trained for Croatian for the grant holder's future benefit.

A grant holder from the Republic of North Macedonia will during his visit in Ljubljana be working with CLASSLA, the CLARIN knowledge centre for South Slavic languages. The objective of the visit is to obtain knowledge about corpora management software in order to be able to learn how to train POS taggers for Macedonian, and to create a corpus, ultimately to be used for corpus-based lexicographic work at the Macedonian Academy of Sciences and Arts. Thus it represents an ideal match between CLASSLA's offer of expertise on language resources and technologies for South Slavic languages, and the ELEXIS objectives of bridging the gap between more advanced and lesser-resourced lexicographic communities.

Through the ELEXIS Travel Grant visits at various ELEXIS hosting institutions that are also part of the CLARIN network, CLARIN was and will continue to be introduced, to a community that without the ELEXIS travel grant opportunity would not have approached a CLARIN centre due to a lack of knowledge. Hence we expect to observe a snowball effect in the future, where ELEXIS grant winners introduce CLARIN within their (national) research community and approach a CLARIN centre thanks to ELEXIS acting as an intermediate. Through the ELEXIS travel grants, the usage of certain CLARIN tools and services, introduced during the ELEXIS research visits, might increase due to additional users and user scenarios.

In addition, ELEXIS aims at establishing interoperability with CLARIN by forming an ELEXIS-CLARIN subgroup[7] to prepare a strategy of integration of ELEXIS services into CLARIN at the end of the project (Summer 2022). We foresee that the subgroup will be formed by members of national CLARIN infrastructures, especially those with CLARIN B centres. Thus, ELEXIS sustainability will be enabled via national consortia to guarantee an afterlife for the efforts made.

5 Conclusion

We have presented the mid-term outcomes of the transnational access initiative of the ELEXIS project where 13 visits to 9 different lexicographical infrastructures in Europe have been completed to date. The grant holders - be they early stage researchers or senior staff - seeked to tailor their research visit

[7] https://elex.is/wp-content/uploads/2019/04/ELEXIS_03_Observer_Session5_ELEXIS_CLARIN-DARIAH.pdf

in a manner that enabled them to gain new knowledge by physically visiting lexicographical milieus with specific expertise in certain topics and technologies that are highly relevant to their research.

The reports of the grant holders clearly show that the travel grants serve several purposes: the above-mentioned gaining of new knowledge and the network building and knowledge exchange are obvious results. For the individual visiting researchers, however, the visits also serve as a career boost either by helping the early stage researchers establish themselves in the field or by leading the more experienced ones towards new fields. The fact that many experienced researchers apply for a grant to deepen their knowledge and gain new expertise shows that the travel grants meet an existing need not covered by other initiatives.

ELEXIS has a network of 50 observer institutions[8] that benefit from early access to newly developed tools and services, as well as to activities aimed at improving and enriching their own lexicographic data. In the upcoming calls, we expect to receive more applications from these observer institutions. Furthermore, the observing institutions (with lexicographic data) will be invited to join in the network of ELEXIS infrastructures who host the travel grants. They will not receive compensation for work at the institution but visitors will be compensated in the same manner as when visiting existing infrastructures.

Most presumably, the grant visits of the second half of the project will follow the same line as the previous ones. However, we expect to see an increased interest in the integrative use of the lexicographical tools, methodologies and resources that are just currently being developed and made available through ELEXIS, e.g. the automated linking tool NAISC (McCrae, 2018) and Lexonomy, an open-source platform for writing and publishing dictionaries (Měchura, 2017), which several grant holders already report having worked with.

Concludingly, the ELEXIS travel grants are an opportunity for the CLARIN infrastructure to get known in a community that without ELEXIS would not approach a CLARIN centre. When the project ends we aim at establishing hopefully full interoperability with existing CLARIN centres, enabled via national consortia.

References

McCrae, J. & Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies. 18*. 109-123. 10.2478/cait-2018-0010.

Měchura, M. B. (2017) 'Introducing Lexonomy: an open-source dictionary writing and publishing system'. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference,* 19-21 September 2017, Leiden, The Netherlands.

Olsen, S. & Pedersen, B.S. (2020). D.9.2 *Report on trans-national access – year 2.* Available at: https://elex.is/wp-content/uploads/2020/07/ELEXIS_D9_2_Report-on-TNA-2.pdf

Woldrich, A. and Wissik, T. (2019). *D7.5 First Year Communication and Dissemination Report and Updated Communication Plan.* Available at https://elex.is/wp-content/uploads/2019/03/ELEXIS_D7_5_First_year_dissemination_and_communication_report_and_updated_communication_plan.pdf

Woldrich, A. and Wissik, T. (2020). *D7.6 Second Year Communication and Dissemination Report.* Available at https://elex.is/wp-content/uploads/2020/05/ELEXIS_D7_6_Second_Year_Dissemination_and_Communication_Report.pdf

[8] https://elex.is/observers/

PoetryLab as Infrastructure for the Analysis of Spanish Poetry

Javier de la Rosa
LINHD
UNED
Madrid, Spain
versae@linhd.uned.es

Álvaro Pérez
LINHD
UNED
Madrid, Spain
alvaro.perez@linhd.uned.es

Laura Hernández
LINHD
UNED
Madrid, Spain
laura.hernandez@linhd.uned.es

Aitor Díaz
Control and Communication Systems
UNED
Madrid, Spain
adiazm@scc.uned.es

Salvador Ros
Control and Communication Systems
UNED
Madrid, Spain
sros@scc.uned.es

Elena González-Blanco
School of Human Sciences and Technology
IE University
Madrid, Spain
egonzalezblanco@faculty.ie.edu

Abstract

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

1 Introduction

The main goal of the ERC-funded POSTDATA Project (Ros and González-Blanco, 2018; Curado Malta and González-Blanco, 2016)[1] was to formalize a network of ontologies capable of expressing any poetic expression and its analysis at the European level, thus enabling scholars all over Europe to interchange their data using Linked Open Data. However, varied research interests result in corpora that might not share the same facets of an analysis. To alleviate this concern and foster the completeness of the interchanged corpora, our team set to build a software toolkit to assist in the analysis of poetry. This paper details the first iteration of the PoetryLab, an extensible open source toolkit for syllabification, scansion,

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

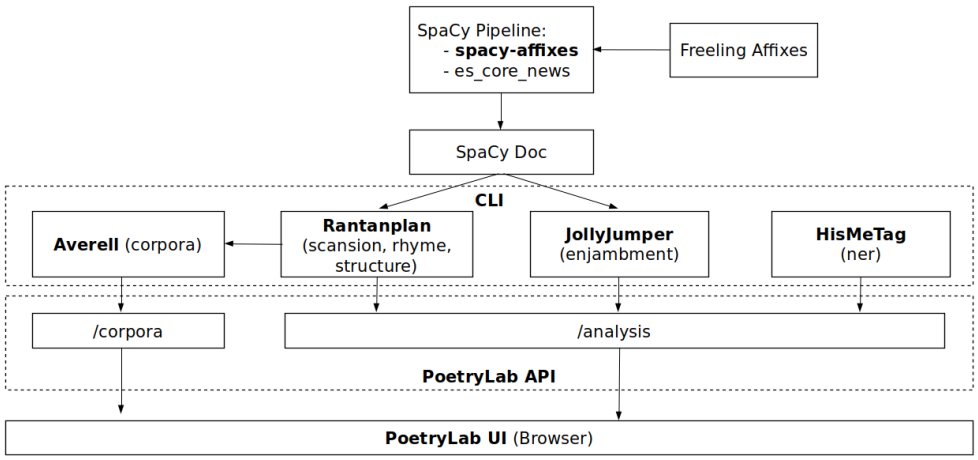


Figure 1: General architecture of the PoetryLab.

enjambment detection, rhyme detection, and historical named entity recognition for Spanish poetry, that achieves state of the art performance in the tasks for which reproducible alternatives exist.

2 PoetryLab

Despite a long and rich tradition (Bello, 1859; Navarro Tomás, 1991; Caparrós, 2014), not many computational tools have been created to assist scholars in the annotation and analysis of Spanish poetry. With ever increasing corpora sizes and the popularization of distant reading techniques (Moretti, 2013; Jockers, 2013), the possibility of automating part of the analysis became very appealing. Although solutions exist, they are either incomplete, e.g., scansion of fixed-metre poetry (Agirrezabal et al., 2016; Navarro-Colorado, 2017; Gervas, 2000; Agirrezabal et al., 2017), not applicable to Spanish (Agirrezabal et al., 2017; Hartman, 2005), or not open or reproducible (Gervas, 2000). Moreover, disparate input and output formats, operating system requirements and dependencies, and the lack of interoperability between software packages, further complicated the limited ecosystem of tools to analyze Spanish poetry. These limitations guided the design of the PoetryLab as a two layer system: a REST API that connects the different tools, and a consumer web-based UI that exposes the functionality to non-experts users. Independently, all tools are released as Python packages with their own command line interface applications (where appropriate), and are ready to produce RDF triples compliant with the POSTDATA Project network of ontologies (Ros and González-Blanco, 2018; Bermúdez-Sabel et al., 2017). Figure 1 shows a diagram of the general architecture of the system.

This granular design allows for each component of the PoetryLab to be used and deployed as a Docker image, which makes managing the different tools lifecycle and versioning a less problematic issue. We tested this approach by using ouroboros[2], a service to automatically update running docker containers with the newest available image, and the demo site of the PoetryLab has been running without major incidents over a year now[3]. We feel hosting the PoetryLab as one of the tools in the catalog of software tools available in CLARIN would be a great addition, since it requires little effort to setup and the maintenance of the different tools is deferred to their own maintainers, as it usually happens in the Open Source ecosystem, making it easy for hot-replacement when new versions become available. Moreover, the use of Docker containers as a deployment strategy and the fact that the tools are stateless, allows the

[2]See `https://github.com/pyouroboros/ouroboros`
[3]See `http://postdata.uned.es/poetrylab`

use of lambda architectures to minimize the running costs.

2.1 PoetryLab API

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

2.1.1 Endpoint `/analysis`

The first endpoint of the PoetryLab API, `/analysis`, leverages three tools to perform several aspects of the analysis of a poem: scansion and rhyme identification, enjambment detection, and named entity recognition. First, built on top of the industrial-strength natural language processing framework spaCy (Honnibal and Montani, 2017), two Python packages perform scansion and rhyme analysis, and enjambment detection, namely, Rantaplan and JollyJumper[5]. AnCora (Taulé et al., 2008), the corpus spaCy is trained on for Spanish, splits most affixes thus losing the multi-token word information and causing some failures in the part of speech tags it produces. To circumvent this limitation and to ensure clitics were properly handled, we integrated Freeling's affixes rules via a custom built pipeline for spaCy. The resulting package, spacy-affixes[6], splits words with affixes so spaCy can handle their part of speech correctly (Padró and Stanilovsky, 2012). Getting this information right was crucial to identify the stress of some monosyllabic and bisyllabic words, and to find a special kind of enjambment called sirrematic, in which a grammatical unit is divided in two lines (see Table 1 for a summary of the performance of our scansion system). The outputs of these two tools are then transformed to accommodate to the definitions given in the network of ontologies developed within the POSTDATA Project.

Method	**Accuracy (%)**
(Gervas, 2000)[7]	70.88
(Navarro-Colorado, 2017)	94.44
(Agirrezabal et al., 2017)	90.84
Rantanplan (ours)	**96.23**

Table 1: Percentages of accuracy on Navarro-Colorado's fixed-metre 1,400 verses corpus. Best scores in bold.

Lastly, the PoetryLab API provides a pluggable architecture that allows for the integration of external packages developed in languages other than Python. This is the case for our named entity recognition system, HisMeTag (Platas et al., 2017), developed in Java and connected to the PoetryLab API through an internal REST API exposed via Docker. The only requirement is to consume raw plain text and produce both a JSON output and RDF triples compliant with the POSTDATA Project network of ontologies.

2.1.2 Endpoint `/corpora`

The second available endpoint, `/corpora`, aims to facilitate working with existing repositories of annotated poetry. Averell[8], the tool that handles the corpora, is able to download an annotated corpus and reconcile different TEI entities to provide a unified JSON output and RDF triples at the desired granularity. That is, for their investigations some researchers might need the entire poem, poems split line by line, or even word by word if that is available. Averell allows specifying the granularity of the final generated dataset, which is a combined JSON or RDF file containing all the entities in the selected corpora.

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

[5] See `https://github.com/linhd-postdata/rantanplan/` and `https://github.com/linhd-postdata/jollyjumper`

[6] See `https://github.com/linhd-postdata/spacy-affixes/`

[7] Gervás' method was only evaluated on 9,643 verses of the 10,000 verses corpus.

[8] See `https://github.com/linhd-postdata/averell/`

Name	Size	Docs	Words	Granularity	License
Disco V2	22M	4088	381539	stanza, line	CC-BY
Disco V3	28M	4080	377978	stanza, line	CC-BY
Sonetos Siglo de Oro	6.8M	5078	466012	stanza, line	CC-BY-NC 4.0
ADSO 100 poems corpus	128K	100	9208	stanza, line	CC-BY-NC 4.0
Poesía Lírica Castellana del Siglo de Oro	3.8M	475	299402	stanza, line, word, syllable	CC-BY-NC 4.0
Eighteenth Century Poetry Archive	2400M	3084	2063668	stanza, line, word	CC-BY-NC 4.0

Table 2: Available corpora in Averell (Ruiz Fabo et al., 2018; Navarro-Colorado et al., 2016; Huber, 2018)

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

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

2.2 PoetryLab UI

The PoetryLab API is then used to provide with functionality to a React-based web interface that non-technical scholars can use to interact with the packages in a graphical way (see Figure 2). The frontend gives the option to download the generated data in both JSON and POSTDATA Projet RDF triples formats[9].

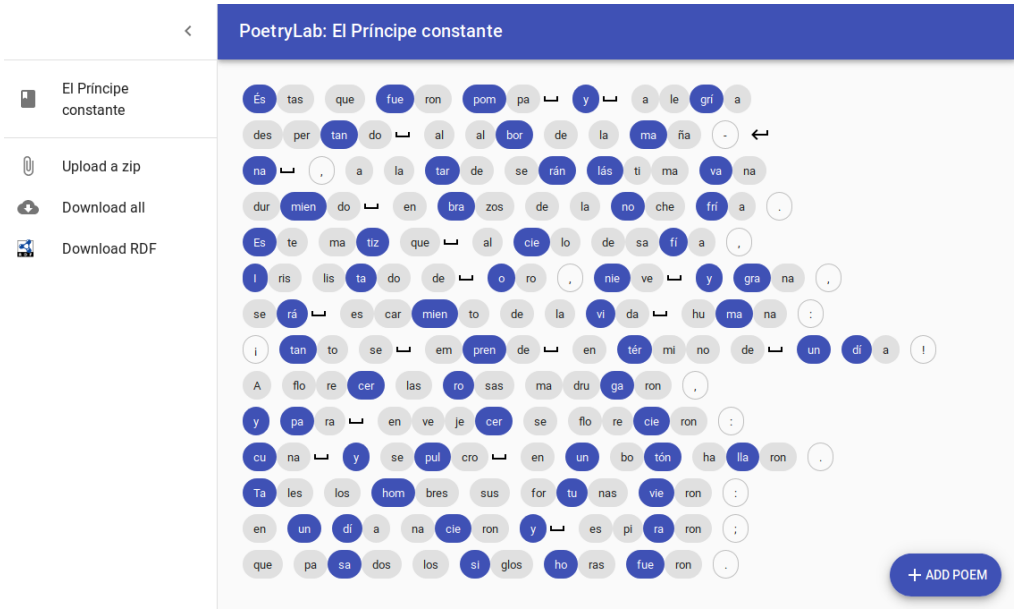


Figure 2: PoetryLab showing stressed syllables (blue), sinalefas (‿) and enjambments (↵).

[9] http://postdata.uned.es/poetrylab

The web interface is run entirely in the browser as a stateless application. However, the collection of analyzed poems are saved to the browser local storage which persist between sessions and restart. Unfortunately, there is no a user management system implemented which would provide with persistent storage in the backend.

3 Conclusion

Although at an early stage, the PoetryLab has proven useful in that it provides an integrated set of tools for Spanish poetry scholars. It also produces machine readable and interoperable data suitable to be ingested into a triple store compliant with the POSTDATA Project network of ontologies. In fact, this approach is already being tested as we export the analysis of poems and feed them into an Omeka that integrates with the POSTDATA Project network of ontologies.

The PoetryLab will be eventually integrated into the larger POSTDATA Project public website, making working with European repositories of poetry a more pleasant task, and assisting whenever possible with the metrical and rhetorical side of the analysis.

Acknowledgements

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

References

Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta, and Mans Hulden. 2016. Zeuscansion: a tool for scansion of english poetry. *Journal of Language Modelling*, 4.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2017. A comparison of feature-based and neural scansion of poetry. *arXiv preprint arXiv:1711.00938*.

Andrés Bello. 1859. *Principios de la ortolojía i métrica de la lengua castellana.*. la Opinión.

Helena Bermúdez-Sabel, Mariana Curado Malta, and Elena Gonzalez-Blanco. 2017. Towards interoperability in the european poetry community: the standardization of philological concepts. In *International Conference on Language, Data and Knowledge*, pages 156–165. Springer.

José Domínguez Caparrós. 2014. Teoría métrica del verso esdrújulo. *Rhythmica: revista española de métrica comparada*, 12:55–96.

Mariana Curado Malta and Elena González-Blanco. 2016. Postdata. towards publishing european poetry as linked open data. In *International Conference on Dublin Core & Metadata Applications*. DCMI.

Pablo Gervas. 2000. A logic programming application for the analysis of spanish verse. In *International Conference on Computational Logic*, pages 1330–1344. Springer.

Charles O Hartman. 2005. The scandroid 1.1. *Software available at http://oak. conncoll. edu/cohar/Programs. htm*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. *Convolutional Neural Networks and Incremental Parsing*, 7.

Alexander Huber. 2018. Eighteenth-century poetry archive.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Franco Moretti. 2013. *Distant reading*. Verso Books.

Borja Navarro-Colorado, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *International Conference on Language Resources and Evaluation*, pages 4360–4364.

Borja Navarro-Colorado. 2017. A metrical scansion system for fixed-metre spanish poetry. *Digital Scholarship in the Humanities*, 33(1):112–127.

Tomás Navarro Tomás. 1991. Métrica española. *Reseña histórica y descriptiva*, 50.

OpenAPI Initiative. 2017. Openapi specification. *Retrieved from GitHub: https://github. com/OAI/OpenAPI-Specification/blob/master/versions/3.0*, 1.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *International Conference on Language Resources and Evaluation*.

M Luisa Díez Platas, Mª de los LLanos Tobarra Abad, Salvador Ros Muñoz, Elena González-Blanco García, Antonio Robles Gómez, Agustín Caminero Herráez, and Gimena del Rio Riande. 2017. Hispanic medieval tagger (hismetag): una aplicación web para el etiquetado de entidades en textos medievales. Zenodo.

Salvador Ros and Elena González-Blanco. 2018. Poetry and digital humanities making interoperability possible in a divided world of digital poetry: Postdata project.(presentation).

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Isabel Martínez Cantón, Elena González-Blanco García, and Borja Navarro-Colorado. 2018. The diachronic spanish sonnet corpus (disco): Tei and linked open data encoding, data distribution and metrical findings.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *International Conference on Language Resources and Evaluation*.

Reproducible Annotation Services for WebLicht

Daniël de Kok and **Neele Falk**
Seminar für Sprachwissenschaft
University of Tübingen, Germany
{daniel.de-kok, neele.falk}@uni-tuebingen.de

Abstract

In this work we report on our experiences with using the Nix (Dolstra, 2006) package manager to build fully reproducible annotation services for the WebLicht (Hinrichs et al., 2010) workflow engine.

1 Introduction

Reproducibility is one of the core foundations of scientific research (Cohen et al., 2018). In the context of automatic annotation tools, such as those provided by WebLicht (Hinrichs et al., 2010), reproducibility means that annotating a particular text at two points in time with a particular model should result in the same annotations. Unfortunately, reusing an annotation tool does not guarantee reproducible results. Both models and software are continuously updated to provide the best results in terms of accuracy. To guarantee reproducible annotations, we need to account for the *software* and its transitive dependencies, the *models* that are used, and circumstantial factors of the *environment* that may influence the behavior of the software.

We will now discuss each of these factors that influence reproduciblity in more detail, before outlining our approach to reproducible services. Throughout this work, we will use the *sticker*[1] annotation tool as a case study. sticker is a part-of-speech tagger, topological field analyzer, and dependency parser that uses word embeddings and deep neural networks. sticker is written in the Rust programming language and uses the Tensorflow C/C++ library for linear algebra operations.

Software Changes between versions of an annotation tool can impact reproducibility. Most trivially, a program could drop compatibility for older models. But even if a program claims to be backwards compatible with models, behaviorial changes of the software may lead to subtle differences in the annotations that are produced. For example, sticker uses the ℓ_2-normalized arithmetic mean of all word embeddings to represent unknown words. However, this was found to be too slow for quantized embeddings. Consequently, for quantized embeddings the representation of unknown words was replaced by the ℓ_2-normalized mean of all subquantizers.[2] This change in representation of unknown tokens can result in different annotations, despite the same model being used.

Changes in transitive dependencies make reproducibility even more difficult to tackle. For instance, Tensorflow, which is a transitive dependency of sticker, produced subtly different outputs for the logistic function depending on whether vector components were handled by a SIMD or scalar code path.[3] When such an issue is fixed, this leads to small numeric differences that can cumulate into different model predictions. Consequently, for reproducibility, the version of the annotation tool, *all* its transitive dependencies, and their build configurations must be fully specified.

[1] https://github.com/stickeritis/sticker
[2] https://github.com/stickeritis/sticker/issues/119
[3] https://github.com/tensorflow/tensorflow/issues/33878

Models In this work, we are interested in reproducibility using existing models, rather than reproducibility in training a model. Reproducibility of models by themselves can be trivially guaranteed by ensuring that a model is bitwise identical.

Environment Even if an annotation tool, its transitive dependencies, and the models are completely specified, there may be environmental factors that impact reproducibility. These environmental factors introduce some form of randomness that make annotation non-deterministic. For example, in some forms of parallelization, a non-deterministic ordering of operations may lead to different rounding errors between runs. Another example is that in unsafe languages, a reliance on uninitialized memory may introduce non-determinism. For the remainder of the paper, we will not discuss such non-determinisms, since they can only be fixed in the upstream software.

In Section 2, we will discuss prior work in reproducible annotation tools and their shortcomings. We will outline our approach to reproducible annotations tools in Section 3. Finally, we will discuss this approach more concretely using sticker as a case study in Section 4.

2 Prior work

2.1 Docker

Most prior work in reproducibility has focused on the use of Docker and particularly so-called Dockerfiles (Cito et al., 2016). Unfortunately, Dockerfiles only provide very weak reproducibility. The prototypical Dockerfile starts along the following lines:

```
FROM ubuntu:xenial
RUN apt-get update
```

The first line specifies the base image that should be used, namely the image tagged *xenial* from the *ubuntu* Docker Hub account. Docker image tags are mutable, so there is no guarantee that exactly the same base image is retrieved between different builds of a Dockerfile. The second line updates the package cache of the APT package manager that is used by Ubuntu. This is often necessary because distributions purge older versions of packages. As packages are updated over time, this will result in different Docker images. The same problems holds for any other step in a Dockerfile that relies on external resources.

Despite the lack of reproducibility in Dockerfiles, the images that are generated by building Dockerfiles provide reproducibility. However, there are two common scenarios that container images cannot accommodate: (1) Making small modifications to an annotation tool that do not impact annotations. This is a typical requirement in infrastructure projects, where e.g. the data exchange format or execution environment can change. (2) Making modifications to an annotation tool for research. In research, we often want to evaluate a change to a system while keeping the rest of the system identical, so that any measured improvements of the system can be attributed to this change.

Another issue with Docker images is that they target a specific container runtime environtment. The same build infrastructure cannot be leveraged to build artifacts for other runtime environments, such as standalone binaries, virtual machine images, system containers (e.g. LXD), or images specific to a cloud provider .

Despite these limitations, Docker containers are a very good mechanism for delivering and running applications. So ideally, a good solution to the reproducibility problem would still be able to produce Docker images.

2.2 Language package managers

Several language package managers, such as Cargo (Rust) and NPM (Node.js) introduce the notion of a *lock file* in which the versions and hashes of transitive package dependencies of a program are specified. This guarantees that the exact dependencies that the developer specified are used. Such lock files are a large step forward in reproducibility compared to package managers that do not specify exact dependencies. However, lock files are limited to specifying dependencies from within the language ecosystem, they do not specify any other dependencies that are used.

3 Our approach

3.1 Introduction

In order to get the reproducibility guarantees sought after in Section 1, we need *hermetic builds*. That is, the model, software, and their dependencies should be fully specified, without any reliance on implicit dependencies. We use Nix (Dolstra, 2006) to realize hermetic builds of annotation services. Nix ensures reproducibility in three ways:

1. Packages (and their transitive dependencies) are defined in the Nix expression language. Nix is a pure, lazy, functional language. Since Nix is a pure language, a package definition will always evaluate to the same so-called *derivation* and any derivations that it depends on.

2. Derivations, which describe how a package is built, are realized in a sandboxed environment without filesystem or network access. Only dependencies that are explicitly specified in the derivation are made available. Sandboxing ensures all build dependencies of a package are fully specified.

3. Each package is realized into a unique path in the *Nix store* and is not globally visible. In typical Unix-like systems, packages are installed into a global namespace such as `/usr` or `/usr/local`. In Nix, each package is installed in its own Nix store path of the form `/nix/store/<hash>-<name>-<version>`, where *hash* is a SHA-256 hash of the derivation. Since such store paths do not have global visibility, any external paths that a package dependends on must be fully qualified. For example, a program that uses the *zlib* shared library will not use a global library dependency such as `libz.so.1`, but will instead use the full path of the *zlib* version it was built against, such as `/nix/store/wfizm...-zlib-1.2.11/lib/libz.so`.

   This approach has two important benefits: it avoids accidental use of another version of a dependency, which may break reproducibility. Secondly, this allows Nix to store multiple versions or build configurations of the same package.[4]

Even though the purity of Nix guarantees that a given package expression will always result in the same derivation, this would not lead to meaningful reproducibility if the sources of the programs and libraries that are built could be replaced. To ensure that external inputs to the build process are always the same, Nix introduces the notion of a *fixed-output derivation*. Fixed-output derivations can be used to fetch external data such as source archives. Fixed-output derivations specificy the SHA-256 hash of the data, so that building a derivation always results in a bitwise identical Nix store path.

3.2 Package ecosystem

Writing Nix derivations for every dependency of a software package would be a large amount of work. The *nixpkgs* repository[5] provides a large, coherent set of derivations that can be built on. As of writing, nixpkgs contains derivations for over 60,000 packages.[6] This includes many packages that are commonly used in research, such as PyTorch, SciPy, or R; but also many domain-specific packages such as spaCy, NLTK, fastText, or Huggingface Transformers from natural language processing.

The complete package set is built several times every day and the results are stored in the nixpkgs *binary cache*. The binary cache is used to avoid that nixpkgs users have to build the packages on their own machine, but also provide long-term availability.[7] Moreover, sources used by nixpkgs derivations are archived by the Software Heritage.[8] An increasing number of derivations can use the Software Heritage archive as a fallback in the rare event that the original upstream source is not available anymore *and* the package was removed from the binary cache.

In the next section, we will describe how we used Nix and nixpkgs to package sticker as a fully reproducible language annotations service.

[4] The use of the derivation hash as part of the Nix store path ensures that every package variation has its own unique path.

[5] `https://github.com/NixOS/nixpkgs`

[6] `https://repology.org/repository/nix_unstable`

[7] The binary cache has never been garbage-collected up to this point: `https://discourse.nixos.org/t/how-long-until-an-output-path-disappears-from-nixpkgs-cache/7606/2`

[8] https://www.tweag.io/blog/2020-06-18-software-heritage/

4 Case study: sticker

Models Since Nix is a small functional language, we can compose more complex functions from basic building blocks. For sticker models, we define the `stickerModel` function which creates a derivation for a model. For example, a German model for part-of-speech tagging can be defined as follows:

```
de-pos-ud = stickerModel {
  modelName = "de-pos-ud";
  version = "20190821";
  sha256 = "163himdn8l...";
  wordEmbeds = fetchEmbeddings {
    name = "de-structgram-20190426-opq";
    sha256 = "0b75bpsrfx...";
  };
 };
```

Here, `stickerModel` is called with an attribute-value set that further specifies the model, in particular the model name and version. In order to fetch the model, `stickerModel` creates a fixed-output derivation using the hash specified in the `sha256` attribute. The model also requires a set of word embeddings, which are specified in the `wordEmbeds` attribute, using the `fetchEmbeddings` function that we have also defined.

Software As described in Section 1, sticker is written in the Rust programming language and uses several Rust and C++ libraries. To build sticker (outside Nix), one would normally use Rust's Cargo build tool. Cargo downloads all the Rust library dependencies of a project, as specified in the project's *lock file*. Afterwards, Cargo will proceed to build the dependencies and finally the project itself.

To build a Nix package of sticker, we cannot run Cargo directly since it requires network access, which is prevented by the Nix sandbox. Instead, we use the `buildRustPackage` function:[9]

```
sticker = buildRustPackage {
  pname = "sticker";
  version = "0.4.0";

  src = fetchFromGitHub { rev = version; sha256 = "..."; ... };
  cargoSha256 = "0v9swm6psa...";

  buildInputs = [ libtensorflow ];
}
```

This function is used as follows. The sources of sticker are specified in the `src` attribute. The `fetchFromGitHub` function creates a fixed-output derivation to retrieve sticker from its GitHub repository. This results in a bitwise identical version of sticker, including its Cargo lock file. The lock file is then used to define another fixed-output derivation for a package that contains the sources of all Rust dependencies of sticker. The hash in the `cargoSha256` attribute is used for this fixed-output derivation. Finally, the dependency on the Tensorflow C++ library is declared through the `buildInputs` attribute. `libtensorflow` references a derivation that specifies how Tensorflow is built.

Docker and virtual machine images For each model, we generate a wrapper derivation that combines the model with sticker. This wrapper provides scripts to start sticker with the particular model.

In turn, from the wrapper derivations we generate derivations that build layered Docker images. The Docker image derivations are created using the `buildLayeredImage`[10] function, which is simply provided with the wrapper derivation and some metadata:

```
dockerImage = buildLayeredImage {
  name = "danieldk/sticker";
  tag = "${modelName}-${version}";
  contents = wrapper;
};
```

[9] This function is provided by the Nix package set: `https://github.com/NixOS/nixpkgs`

[10] `https://grahamc.com/blog/nix-and-layered-docker-images`

In a similar vain, we use a `buildVMImages` function to build virtual machine images for every model. The derivations for the wrapper, Docker image, and virtual machine images are generated by an extended version of the `stickerModel` function that was introduced earlier in this section. The extended function evaluates to an attribute-value set where the `model`, `wrapper`, `dockerImage`, and `vmImages` attributes map to derivations for the model, model wrapper, Docker image, and virtual machine images.

The generated Docker images are deployed in a WebLicht Docker environment to expose sticker through the corresponding WebLicht service.

5 Conclusion and future work

In this work, we have explored using the Nix package manager for building reproducible annotation services. We have found, using sticker as a case-study, that Nix makes it possible to make a fully reproducible annotation service with its models and transitive dependencies. Moreover, we have found that the Nix language allows us to make powerful abstractions that makes adding new models trivial. The Nix derivations are made available through a GitHub repository.[11]

An important open question is how such infrastructure can be integrated more strongly with a workflow engine such as WebLicht. Currently, WebLicht exposes the latest version of a service. To guarantee reproducible annotations, several changes in the workflow engine are necessary. The user should be able to select the version of the service to be used for annotating the text. After the user has specified a valid version, either the corresponding Docker image can be selected, or the Nix machinery can be invoked to build the image first. The versions of the annotation services should be recorded in the WebLicht annotation chain metadata that is stored in the TCF (Heid et al., 2010) output of WebLicht. This will allow other researchers to run identical annotation pipelines in the future.

6 Acknowledgements

We would like to thank Patricia Fischer for her feedback on this work. The financial support of the research reported in this paper is provided by the German Federal Ministry of Education and Research (BMBF), the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK) as part of CLARIN-D.

References

Jürgen Cito, Vincenzo Ferme, and Harald C Gall. 2016. Using Docker containers to improve reproducibility in software and web engineering research. In *International Conference on Web Engineering*, pages 609–612. Springer.

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E. Hunter. 2018. Three Dimensions of Reproducibility in Natural Language Processing. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Eelco Dolstra. 2006. *The purely functional software deployment model*. Ph.D. thesis.

Ulrich Heid, Helmut Schmid, Kerstin Eckart, and Erhard W Hinrichs. 2010. A corpus representation format for linguistic web services: The D-SPIN Text Corpus Format and its relationship with ISO standards. In *LREC*.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-based LRT services for German. In *Proceedings of the ACL 2010 System Demonstrations*, pages 25–29, Uppsala, Sweden, July.

[11] `https://github.com/stickeritis/nix-packages`

The CLARIN-DK Text Tonsorium

Bart Jongejan
Department of Nordic Studies and Linguistics
University of Copenhagen, Denmark
bartj@hum.ku.dk

Abstract

The Text Tonsorium (TT) is a workflow management system (WMS) for Natural Language Processing (NLP). The software implements a design goal that sets it apart from other WMSes: it operates without manually composed workflow designs. The TT invites tool providers to register and integrate their tools, without having to think about the workflows that new tools can become part of. Both input and output of new tools are specified by expressing language, file format, type of content, etc. in terms of an ontology. Likewise, users of the TT define their goal in terms of this ontology and let the TT compute the workflow designs that fulfill that goal. When the user has chosen one of the proposed workflow designs, the TT enacts it with the user's input. This untraditional approach to workflows requires some familiarization. In this paper, we propose possible improvements of the TT that can facilitate its use by the CLARIN community.

1 Introduction

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

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

Traditionally, workflow management systems require (expert) users for the construction of workflow designs. A good example of such a system is WebLicht[1] (Hinrichs et al., 2010). The characteristic that sets the TT apart from traditional WMSes is that workflow designs are computed automatically. Central to this computation is an ontology that is used to describe all data objects involved in workflows. Also tools are described by this ontology, since the TT only needs to know what kind of data enters a tool and what kind of data comes out. A technical description that explains how the TT computes workflow designs is in Jongejan (2016). More about the user perspective of the TT, as conceived during the DK-CLARIN project, is in Offersgaard et al. (2011) and in Jongejan (2013).

The structure of this paper is as follows. Section 2 is about the current state of the TT. Section 3 presents a condensed technical outline of the TT. Section 4 explains how the data is described by metadata and how the tools are described in terms of the metadata characterizing their input and output data. Section 5 is an exposition of our plans for making an interface to the TT that can be easy to use by the larger CLARIN community. Section 6 tells where the TT can be found on the internet. Finally, in Section 7, concluding remarks follow.

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

2 Current State

There are currently over 40 tools[2] integrated in the TT, spanning from tokenization to syntactic parsing and from PDF to text conversion to text to speech transformation. Some tools were developed to address the needs of a single project and later generalized to make them useful for a wider segment of researchers. One such spin-off is that the TT is able to annotate TEI P5 formatted texts with lemmas and part of speech tags.

Many of the tools are multi-lingual. The CST lemmatizer (Jongejan and Dalianis, 2009), for example, lemmatizes 28 languages[3]. The Danish linguistic resources for the CST lemmatizer cover three historical periods: medieval, late modern and contemporary. A few tools, such as the Named Entity Recognizer, work only for Danish.

3 Technical Outline

3.1 Software Components

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

3.2 Tool Integration

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

3.3 Workflow Composition

The TT uses dynamic programming with memoization to compute all workflow designs (directed acyclic graphs) that combine

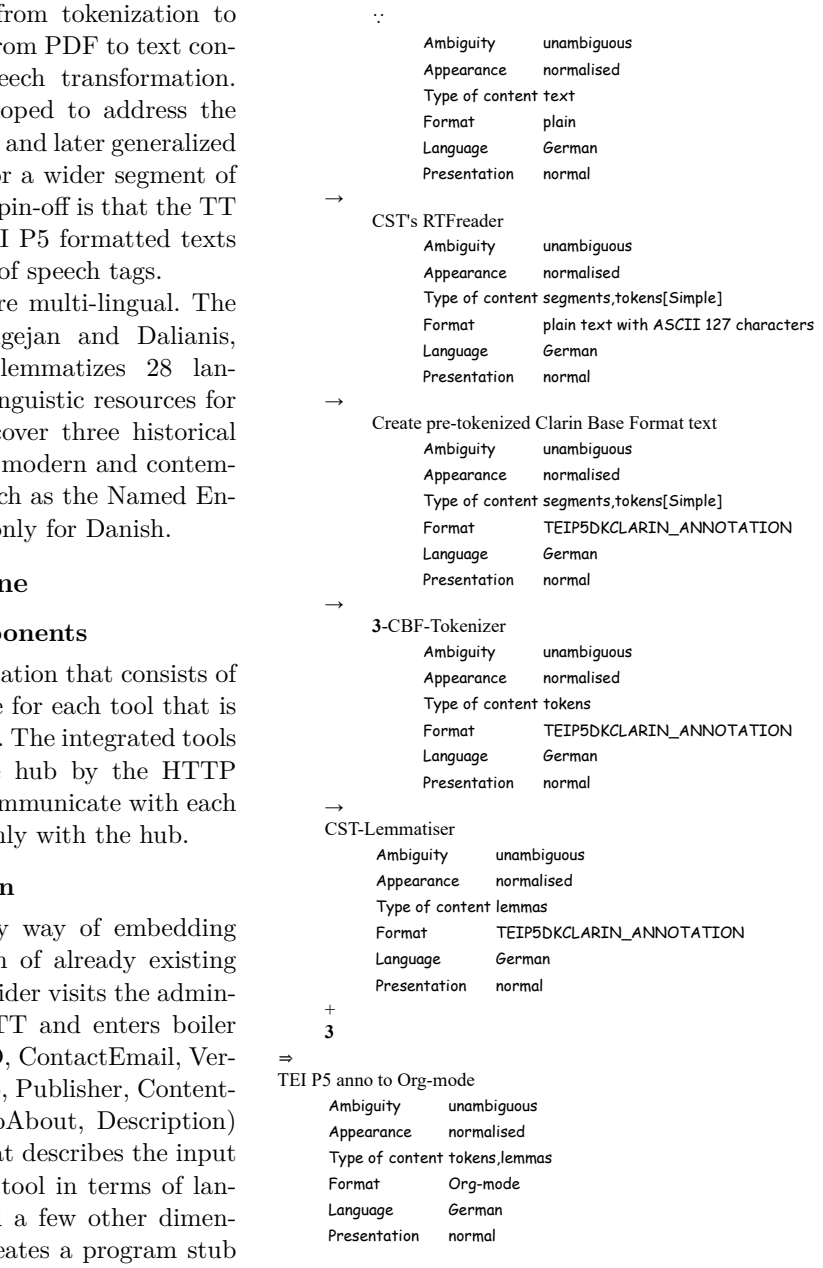


Figure 1: Graph of a workflow design with five tools. The flow is from top to bottom. The output from tools marked with a number is sent to the next tool and to any tool that has the same number mentioned as input. Here, the 3rd tool sends output to the 4th and the 5th tool.

[2]See `ttps://cst.dk/texton/help`

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

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

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

3.4 Workflow Enactment

When the user has chosen one of the proposed candidate workflow designs, the TT enacts that workflow with the data that the user has uploaded as input. The input, intermediary results and final results are temporarily stored on the computer on which the TT runs. The results from each step in a running workflow can be inspected as soon as the step has been executed. The final results can be fetched as a zipped archive.

3.5 Data Formats

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

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

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

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

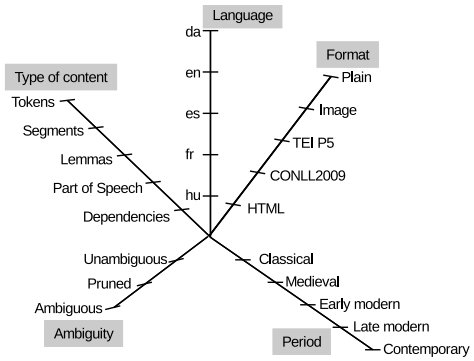
4 Metadata

Five dimensions are used to describe data: the *language*, the *file format*, the *type of content*, the *ambiguity level*, and the *historical period*, see Fig 2.

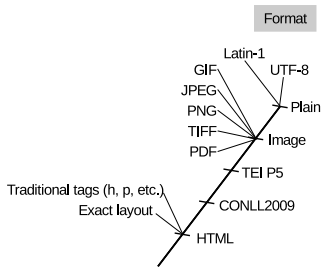
The TT treats all dimensions on an equal footing: it does not care whether a tool transforms data between two *languages* , or between two *file formats*, or between two *types of content*.

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

The listed *types of content* reflect the abilities of the currently integrated tools and of tools that are considered for integration.



(a) Example of independent dimensions.



(b) Examples of subspecification of some values in the file format dimension.

Figure 2: Dimensions and subspecifications of values along the *file format* dimension

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

5 Coming Improvements

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

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

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

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

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

Information about the tool integration process is in (Offersgaard et al., 2011), but the technical details are in project internal documents that we need to rewrite and make publicly available.

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

6 Availability

There is an online version of the TT at the address `https://cst.dk/texton/`. An identical version can be downloaded from GitHub[4]. Some of the tools in the GitHub repository are merely wrappers for open source tools that must be obtained separately, such as LibreOffice (which is used for conversion of office formats to RTF) and the OCR programs Cuneiform and Tesseract.

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

7 Conclusion

We have in a few words described the current state of the Text Tonsorium, a Workflow Manager for NLP tools that automatically computes workflow designs. We have discussed the improvements that we want to implement as part of our efforts to re-launch the Text Tonsorium in CLARIN. We would be pleased to receive more suggestions from users and testers in the CLARIN community.

References

Tomaž Erjavec, Ştefan Bruda, Ivan Derzhanski, Ludmila Dimitrova, Radovan Garabík, Peter Holozan, Nancy Ide, Heiki-Jaan Kaalep, Natalia Kotsyba, Csaba Oravecz, Vladimír Petkevič, Greg Priest-Dorman, Igor Shevchenko, Kiril Simov, Lydia Sinapova, Han Steenwijk, Laszlo Tihanyi, Dan Tufiş, and Jean Véronis. 2010a. MULTEXT-east free lexicons 4.0. Slovenian language resource repository CLARIN.SI.

Tomaž Erjavec, Ivan Derzhanski, Dagmar Divjak, Anna Feldman, Mikhail Kopotev, Natalia Kotsyba, Cvetana Krstev, Aleksandar Petrovski, Behrang QasemiZadeh, Adam Radziszewski, Serge Sharoff, Paul Sokolovsky, Duško Vitas, and Katerina Zdravkova. 2010b. MULTEXT-east non-commercial lexicons 4.0. Slovenian language resource repository CLARIN.SI.

Erhard W. Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. Weblicht: Web-based LRT services for German. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, pages 25–29. The Association for Computer Linguistics.

Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 145–153. Association for Computational Linguistics.

B. Jongejan. 2013. Workflow management in CLARIN-DK. In *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 20*, number 89, pages 11–20. Linköping University Electronic Press; Linköpings universitet.

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

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

[4] `https://github.com/kuhumcst/texton`, `https://github.com/kuhumcst/texton-bin`, `https://github.com/kuhumcst/texton-linguistic-resources` and `https://github.com/kuhumcst/DK-ClarinTools`

Integrating TEITOK and Kontext at LINDAT

Maarten Janssen
Institute of Formal and Applied Lingusitics
Charles University, Czech Republic
janssen@ufal.mff.cuni.cz

Abstract

In this paper we describe how the TEITOK corpus platform was integrated with the Kontext corpus platform at LINDAT to provide document visualization for both existing and future resources at LINDAT. The TEITOK integrations also means LINDAT resources will become available in TEI/XML format, and searchable both using Manatee and CWB.

1 Introduction

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

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

2 Kontext and TEITOK

2.1 TEITOK

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

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

How a TEITOK corpus is exported to CWB is defined in the corpus settings - a central configuration file for the corpus in TEITOK that defines, apart from the CWB export, a myriad of other things like which token attributes should be editable, which items should appear in the menu bar, what the fixed

fields in the teiHeader are, etc. The export settings are explicitly not kept in the teiHeader, since they belong to the corpus, and not to the XML files; and also because the same XML file is sometimes used in multiple corpus projects, each with its own settings. Since TEI files are edited in TEITOK, keeping copies of the same file with a different header would lead to inconsistencies.

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

2.2 Combining TEITOK and Kontext

As mentioned before, the decision was made at LINDAT to adopt TEITOK for document visualization. Kontext and TEITOK are comparable frameworks in the sense that both are online platforms to search corpora using the Corpus Query Language (CQL). And intuitively, it would seem to be the ideal solution to fully integrate the two platforms in a single platform by either incorporating the relevant parts of TEITOK into Kontext, or by rewriting TEITOK to use the search engine using in Kontext, Manatee (Rychlý, 2007), as a backbone instead of CWB. But the problem with both these options is that they constitute a major overhaul of either TEITOK or Kontext, neither of which is used exclusively for LINDAT, but used in a growing number of projects around the world. And a major rewrite would inevitably lead to compatibility problems at other projects using these tools. The only feasible option would be to drop Kontext altogether in favour of TEITOK, but not only are most users of LINDAT more used to the Kontext interface than they are to their TEITOK counterparts, but also there are various projects at LINDAT that rely on the Kontext interface.

Therefore, a more modest integration was selected in which the two platforms are kept as independent interfaces, with links leading from one to the other. That set-up has the added advantage that it allows users that are more familiar with the CWB flavour of CQL than they are with the Kontext implementation have the option to use that by doing their searches using TEITOK rather than Kontext.

TEITOK and Kontext use a very similar file structure, and an identical token numbering: every token is numbered sequentially, and *sattributes* (structural markup) do not take up corpus positions. The naming and encoding of the files is slightly different, meaning the files of one system cannot directly be used in the other, but the same corpus position will point to the same token in the two systems. And this forms the basis for the integration of the two platforms. So the same corpus is indexed in both, and corpus positions can be exchanged. In principle, compiling the same corpus in both systems is trivial since both build their corpora from VRT files. However, as mentioned earlier, TEITOK writes CWB files directly and does not use VRT as an intermediate format. Therefore, a more indirect route was chosen: TEITOK is used to directly create a CWB corpus from TEI/XML files, using the standard TEITOK set-up. Once complete, the CWB corpus is exported to VRT using the CWB tools, after which it is loaded into Manatee. The registry file for both platforms is written by TEITOK using the aforementioned corpus settings.

In order to link the two platforms, a small addition was made to both platforms: in Kontext an option was added to allow the context of a token to be provided by an external REST service. And in TEITOK a module was added that can render the XML context of any corpus token as an HTML page. Combining these two modules makes it possible to click on a token in the Kontext search result, and in the pop-up window that shows the context see the TEITOK rendering of the original XML with all its attributes. The fragment also comes with a link to the visualization of the full document in TEITOK. An example

of a TEITOK context from a Czech text from Skript 2015[1] is given in figure 1, where the word *cvičího* (practicing) is deleted in the original and hence not present in the CWB or Manatee corpus, as can be seen in the KWIC line.

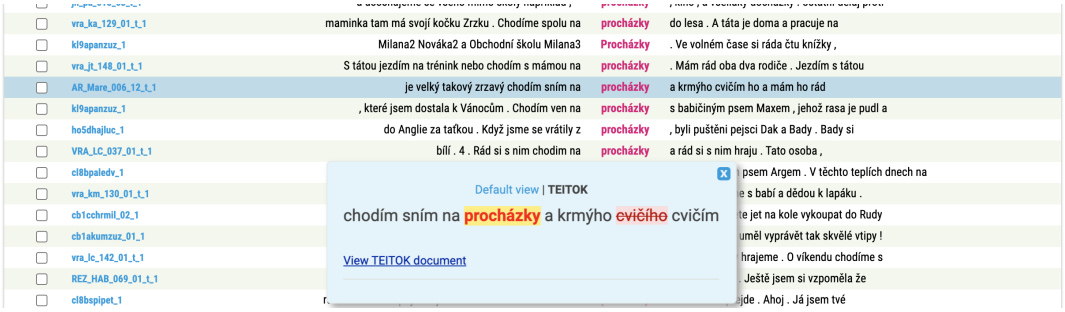


Figure 1: Example of a TEITOK context in Kontext

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

3 Adding resources

Having an integration between the two platforms is not sufficient for an infrastructure: it is also needed to get corpora into the hybrid system. When adding corpora to the integrated TEITOK/Kontext infrastructure, one has to distinguish between corpora that are already in Kontext, existing corpora that are not yet in Kontext, and newly planned corpora.

The corpora that are already in Kontext can be converted automatically, although they have to be regenerated in order to create the byte-offset files used by TEITOK. The existing corpus is exported to VRT, from which it is converted to a collection of TEI/XML files, one for each document. Some corpus specific action is needed, for instance, it is necessary to indicate the correct XPath according to TEI for each of the metadata in the corpus in order to end up with correct TEI/XML documents; but once these settings are provided, the process is fully automatic.

For the class of existing corpora that are not yet in Kontext, TEI/XML files are generated from the raw data (in whichever format the corpus came in), keeping as much of the original information as possible. In order to facilitate this, we are working on a set of conversion tools from popular corpus formats including ELAN, FoLiA, PagesXML, etc. Even for some of the corpora that were already in Kontext, the corpus was recreated from source, since some of the information was lost in the conversion. An example of this is the Prague Dependency Treebank (Hajič, 2004), where the original PML files contain more morphological information than the Kontext corpus that was previously created from it. The richer TEI/XML structure of TEITOK makes it possible to incorporate all this information in the hybrid framework.

And finally, there are those corpora that are still under development or planned for the future. For such corpora, the various options provided by TEI/XML and TEITOK make it possible to encode richer information in the corpus than a traditional VRT corpus would allow. And example of this is the upcoming ParCzech corpus (Hladká et al., 2020) which was designed from the start as a spoken corpus in TEITOK,

[1] http://lindat.mff.cuni.cz/services/teitok/skript2015/

ending up as a searchable Kontext corpus using the TEITOK/Kontext infrastucture. In some cases, existing corpora are being replanned to make use of the new options. An example of that is CzeSL (Štindlová et al., 2012), a learner corpus where not all the information present in the source material was encoded, but which is now in part being redone to include the deletions, corrections, normalizations, etc that were previously impossible to include.

4 Conclusion

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

It is our hope that the combined workflow will attract more corpora into the LINDAT infrastructure that would otherwise not have been made available as a CLARIN resource, with the preliminary indications looking promising in newly established collaborations towards this end. The combined TEITOK/Kontext workflow provides a lot of potential for the future, and the fact that TEITOK is now maintained at LINDAT makes it much easier to expand the framework to account for newly arising demands. One area where we are currently trying to make improvements is the field of aligned corpora, where TEITOK allows for the option to align corpora at multiple levels.

References

CLUL. 2014. P.S. Post Scriptum. arquivo digital de escrita quotidiana em portugal e espanha na Época moderna.

Jan Hajič. 2004. Complex Corpus Annotation: The Prague Dependency Treebank. Bratislava, Slovakia. Jazykovedný ústav Ľ. Štúra, SAV.

Barbora Hladká, Matyáš Kopp, and Pavel Straňák. 2020. ParCzech PS7 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 4037–4043.

Maarten Janssen. 2018. Adding words to manuscripts: From pagesxml to teitok. In Eva Méndez, Fabio Crestani, Cristina Ribeiro, Gabriel David, and João Correia Lopes, editors, *Digital Libraries for Open Knowledge*, pages 152–157, Cham. Springer International Publishing.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Miguel Oliviera Jr. 2016. NURC Digital: Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nurc). *CHIMERA: Romance Corpora and Linguistic Studies*, 3(2):149–174.

Pavel Rychlý. 2007. Manatee/Bonito - a modular corpus manager. In *RASLAN*.

Barbora Štindlová, Svatava Škodová, Jirka Hana, and Alexandr Rosen. 2012. CzeSL – an error tagged corpus of czech as a second language. pages 21–32, 01.

CLARINO+ Optimization of Wittgenstein Research Tools

Alois Pichler
Wittgenstein Archives at the University of Bergen
Bergen, Norway
alois.pichler@uib.no

Abstract

The Wittgenstein Archives at the University of Bergen (WAB) offer specialized tools for research access to its Wittgenstein resources which however are in need for an upgrade to better serve user requirements. The paper discusses this need along some selected exemplary features of two such tools: Interactive Dynamic Presentation (IDP) of Wittgenstein's philosophical Nachlass and Semantic Faceted Search and Browsing (SFB) of Wittgenstein metadata. The tasks of extending and better adapting these two tools to user requirements shall be carried out within the Norwegian CLARINO+ project.

1 Data and metadata for Wittgenstein research

During his lifetime, the Austrian-British philosopher Ludwig Wittgenstein (1889-1951) published only one philosophical book, the *Logisch-philosophische Abhandlung* / *Tractatus logico-philosophicus* (1st ed. 1921/22), and a *Dictionary for Elementary Schools* (1st ed. 1926). However, on his death in 1951, he left behind a significant 20,000 page corpus of unpublished philosophical notebooks, manuscripts, typescripts and dictations. This oeuvre, called "Wittgenstein's Nachlass" or "the Wittgenstein papers" (von Wright, 1969), was brought to the wider public through posthumous book publications such as *Philosophical Investigations* (1st ed. 1953) and *Culture and Value* (1st ed. 1977).

The practice of bringing the Nachlass to modern readers through digital editing, hereby creating new access and research possibilities, reached its first milestone in 1998 with Vol. 1 of the Bergen CD-ROM edition *Wittgenstein's Nachlass: The Bergen Electronic Edition* (Wittgenstein, 2000), edited by the Wittgenstein Archives at the University of Bergen (WAB, http://wab.uib.no/). Since its establishment in 1990, WAB has worked towards providing digital data and metadata for conducting Wittgenstein Nachlass research (Huitfeldt, 2006). This includes the creation of machine-readable transcriptions of the Nachlass with specialized markup. These transcriptions are today accessible as HTML outputs through "interactive dynamic presentation" interfaces (IDP, see Pichler and Bruvik, 2014) on WAB's "Nachlass transcriptions" site http://wittgensteinonline.no/ (Wittgenstein, 2016-). Along with high quality Nachlass facsimiles, they are also increasingly available as HTML outputs in WAB's *Bergen Nachlass Edition* on Wittgenstein Source (Wittgenstein, 2015-). In addition, WAB is working on the implementation of semantic web methods and technology for Wittgenstein research and offers free download of a continuously growing Wittgenstein ontology (see Pichler and Zöllner-Weber, 2012) in OWL (RDF) format from its website, as well as an ontology explorer for semantic faceted search and browsing (SFB, http://wab.uib.no/sfb) of the ontology.

However, WAB's transcriptions and facsimiles of the Nachlass, its metadata for the Nachlass and Wittgenstein research more generally, as well as the tools for making all these accessible are in

need of upgrades for the research community to be able to make the most of them. In this paper I shall focus on some selected exemplary aspects of the required optimization of the IDP and SFB tools. The tasks of extending and better adapting these two tools to user requirements shall be carried out within the Norwegian CLARINO+ project.

2 Interactive Dynamic Presentation (IDP) of Wittgenstein's Nachlass

Digital editions offer significant advantages over print editions in that they allow dynamic and user-tailored access to the material edited. WAB's transcriptions of the Wittgenstein Nachlass contain XML TEI (P5) markup with detailed philological and semantic information about each of the Nachlass items, pages, remarks, sentences, formulas, drawings, words, letters, and characters. The IDP tool to access these transcriptions is built on XML technologies (using Xalan XSLT), HTML and PHP, and permits the user to produce HTML clean copy, "linear" outputs of the Nachlass as well as "diplomatic" outputs recording all deletions, insertions, overwritings etc.[2]

At first, most users simply want to access, read and search the Nachlass through these two primary presentation formats. But they soon discover that they still need other types of versions: for example, a version that in contrast to the diplomatic output omits those deletions where the parts deleted don't fit syntactically into the context and thus "disturb" the reading; or a version that in contrast to the standard linear rendering still retains deleted parts (but shows that they are deleted by Wittgenstein) the inclusion of which may help understand the text; or a version which in contrast to the linear version offers full up-to-date standardization of orthography and punctuation. Yet other requirements include the possibility to filter and arrange the Nachlass corpus according to the editorial marks which Wittgenstein often assigns to his remarks, while at the same time also retaining the possibility to include or omit the marks themselves in the editorial output. This feature permits, for example, extraction of all remarks and only the remarks which are marked by Wittgenstein with a slash, or an asterisk, a backslash, etc., or a specific combination of them, and can help the user see the genetic processes behind the Nachlass or recognize thematic groups. Still another required feature is the possibility to interactively remove in the transcription of a typescript all handwritten revisions and thus to produce a version of the typescript in its purely machine-typed form. This function comes handy when one wants to compare vocabulary and concepts before and after the revision and was indirectly already asked for almost fifty years ago when A. Kenny (1976) wished for an edition of the Big Typescript "as it stood" (i.e. before Wittgenstein's revisions of it). Finally, yet other features required and crucial to Wittgenstein Nachlass research include the possibility to organize the remarks of an item chronologically rather than in their physical order (which often deviates from the chronological one), and the possibility to arrange Nachlass texts according to Wittgenstein's editorial numbers that he uses to group his remarks according to topic.

The possibility to control WAB Nachlass transcription outputs through IDP tool parameters such as the ones mentioned above is valuable to Wittgenstein research, and many of the features described are prepared for in WAB's transcriptions through specific XML TEI encoding. Chronological sorting, for example, can be implemented on the basis of the fact that WAB's transcription records contain a date for each single Nachlass remark; omission of handwritten revision in typescripts can be achieved thanks to WAB's explicit encoding of handwriting in typescripts; filtering and sorting of the Nachlass texts according to Wittgenstein's editorial numbers could be put to practice thanks to the specific encoding WAB uses for them, etc. etc. But to put all relevant encodings to work and to offer them for IDP toggling demands a substantial programming investment. While some of the features required and mentioned above already work, many do not yet work flawlessly or on all required levels. Chronological sorting, for example, currently only works on the level of single items, although it is precisely at higher levels that it may be needed the most, for example, in the chronological arrangement of Nachlass item *groups* or even the *entire* Nachlass corpus. In brief, the IDP tool currently manages to offer access to (1) only a fraction of the encoding, (2) only a fraction of combinatorial possibilities of the encoding, (3) only a fraction of the presentation, sorting and filtering possibilities, and (4) in all these three fields it is susceptible to errors due to undesired interference. It is in fact a major challenge to provide for presentation modes

[2] For a more thorough explanation of the terms "diplomatic" and "linear", see Pierazzo (2009).

that work in tandem and that can be accumulated rather than interfering negatively with each other. This factor results in limitations such as the following: with regard to (1), users cannot yet filter the transcriptions for insertions of a specific subtype; with regard to (2), users are not yet able to combine filtering of insertions with filtering of the encoding of text alternatives; and with regard to (3), it is not yet possible to render the type of insertions selected in ways other than what is set by WAB as the default for the IDP site. Moreover, with regard to (2), it is for the user currently not possible to combine a marking of Wittgenstein's text alternatives with a diplomatic rendering, or with the inclusion / exclusion of his own markers for text alternative, or with a toggling of including / excluding the alternatives discarded by him.

Users with XML programming competence may be able to respond to all such needs by directly processing and querying the XML transcription file itself. This is one of the reasons why also within the framework of the CLARINO+ project WAB will deposit its transcription corpus in the CLARINO Bergen Repository. This task includes generation of CLARIN CMDI-conformant metadata as well as designing licenses for the use of both the transcriptions and the metadata offered. Previously within the frameworks of the Cost A32 and Discovery projects, WAB has already made available XML transcription samples of 5000 Nachlass pages under the CC BY-NC 4.0 license.[3]

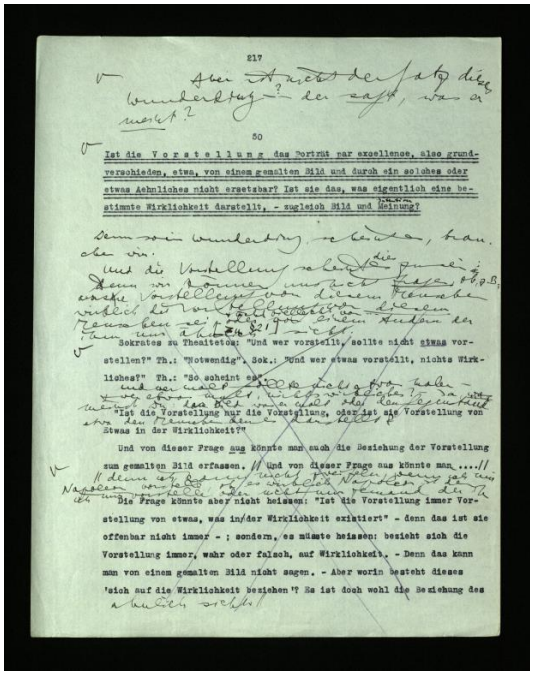


Figure 1: Facsimile of Wittgenstein Nachlass Ts-213,217r, reproduced with the kind permission of The Master and Fellows of Trinity College Cambridge and the University of Bergen. CC BY-NC 4.0. `http://www.wittgensteinsource.org/Ts-213,217r_f`

The page displays (most relevant for IDP) handwritten revisions of the typescript incl. deletions, insertions, markings, variant writing and frequent use of the editorial mark "v". The page also contains (most relevant for SFB) at remark Ts-213,217r[3], "Sokrates zu Theaitetos …", a reference to Plato, as well as the internal reference "[Zu § 21]". Transcriptions of this remark are made available by WAB at `http://www.wittgensteinsource.org/Ts-213,217r[3]_d` *(diplomatic version),* `http://www.wittgensteinsource.org/Ts-213,217r[3]_n` *(linear version) and (interactive dynamic presentation)* `http://wittgensteinonline.no/`.

3 Semantic Faceted Search and Browsing (SFB) of Wittgenstein metadata

WAB's reference system assigns a unique identifier to each remark in the Nachlass, called "siglum". The siglum provides a URL for each single Nachlass component and makes up the backbone of the Wittgenstein ontology and the SFB site that offers semantic faceted search and browsing of WAB's metadata for the Wittgenstein domain. The SFB tool is built on the University of Bergen Library's search infrastructure, which involves technologies such as Elasticsearch,

[3] See `http://wab.uib.no/cost-a32_xml/`; for the aforementioned two projects see `http://wab.uib.no/wab_R&D.page`. It must be noted that the entire Wittgenstein Nachlass is made available by WAB as open access in the sense of *free*, but only the mentioned 5000 pages are currently open access also in the sense of *libre* open access (for the distinction see Suber, 2003). For the licenses for all WAB resources currently offered to the public, see Pichler (2019).

Apache Jena and Angular framework.[4] Today, SFB already permits search and browsing of Nachlass remarks along a number of facets, incl. reference to a person, reference to a work, its dating and its relation to "published works", and displays the resulting remark hit along with a link to the corresponding facsimile in Wittgenstein (2015-). However, while much more metadata are recorded in the transcription or by stand-off markup, they need first to be modeled and ingested into the tool. Examples are information about a remark's genetic path(s), its place of origin in the Nachlass corpus, references to places, events and other named entities, similarity to other remarks (see Ullrich, 2019), adherence to text type and genre (philosophical remark, preface, motto, dedication, instruction, aphorism, diary entry, autobiographical remark, mathematical-logical notation, graphic etc.), adherence to Nachlass group (notebook, loose sheet, "Zettel", ledger, typescript, dictation etc.), work status (first draft, elaborated version, final work etc.), script type (short hand, secret code etc.), the language the remark is written in, the writing material (pencil, ink etc.), research literature referring to it, and other.

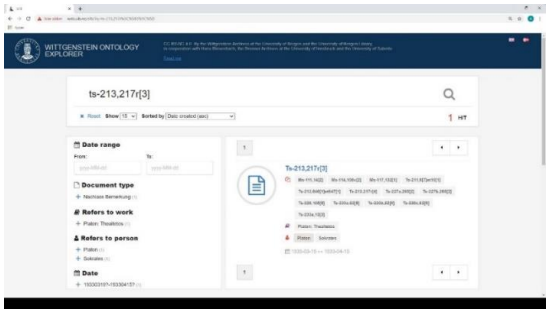


Figure 2: Screenshot of the SFB hit for Wittgenstein Nachlass remark Ts-213,217r[3], available at `http://wab.uib.no/sfb/?q=ts-213,217r%5C%5B3%5C%5D`

An important contribution to the SFB service will be the inclusion of a digital Wittgenstein lexicon (see Röhrer, 2019), as an outcome of the cooperation between WAB and the Centrum für Informations- und Sprachverarbeitung (CIS) at the Ludwig Maximilians Universität München on the search tool WiTTFind (`http://wittfind.cis.lmu.de/`) – WAB contributing its facsimiles and encoded XML transcriptions of the Wittgenstein Nachlass as well as XSLT stylesheets for their processing, and CIS providing programming and computational linguistics personnel resources as well as a grammatically encoded digital lexicon of the German language (see Hadersbeck, Pichler et al., 2016). WiTTFind offers lemmatized online text search access to the entire Nachlass, displays each sentence containing any grammatical form of the word searched for within the context of the larger remark, and additionally highlights the hit in the corresponding facsimile of the remark. WiTTFind continues WAB's siglum reference system for the Nachlass even down to sentence level. Recently, the cooperation project has also embarked on a word tokenizer based on WAB's XML transcriptions of the Nachlass. Implementing WiTTfind in the SFB tool will permit simultaneous and combined SFB of *both metadata and text data*. Researchers interested in the genesis of Wittgenstein's philosophy, for example, may want to know when Wittgenstein started to replace the expression "calculus" with the expression "game", and whether this development can be linked to any other development, e.g. increased reference to works of others, other changes in vocabulary, developments in letter correspondence, meetings and discussions with friends and colleagues, etc. An integration of the SFB and WiTTFind tools will bring us closer to seeing all the connections between the Nachlass' remarks and contents. WiTTFind is a fine example of the added value created by making one's data available for research and reuse by others, and CLARINO+ is a fine example of capitalizing this value further by implementing its outcomes into the CLARIN confederation of language and text resources.

CLARINO+ will both integrate the WiTTFind lexicon into CLARIN and improve the SFB tool itself. Outstanding tasks include correcting errors and deficiencies in the overall browsing and combinatorial setup and adding and organizing facets still lacking; one example is chronological sorting of a remark's variants, which currently are only displayable in alphanumeric order (for an

[4] See `http://marcus.uib.no`, `https://www.elastic.co`, `https://jena.apache.org/` and `https://angular.io/`.

example see Figure 2). The upgrade will require improvements of the user interface, including addition of display labels for creation dates along with search results. A highly desired addendum is the possibility to view the remark hit resulting from one's searching and browsing along with a linear or diplomatic transcription of the remark; currently only the remark's siglum along with a link to the corresponding facsimile is displayed.

4 Conclusion

Although at present WAB, along with its IDP and SFB tools, already enjoys a large number of international users[5], it is only when deficiencies such as the ones described above are corrected and further requirements and desiderata fulfilled, that researchers will be able to take full advantage of WAB's resources. Only then will users be equipped to fully exploit the multifaceted interrelations between and within Wittgenstein data and metadata provided by WAB for the community's research questions. At the same time, it is also then that the deep issues about the relation between on the one hand the contents and forms of Wittgenstein's philosophy and work, and on the other hand their interpretation and application, can properly begin to play out in sufficiently complex formats via interactive digital media.

References

Max Hadersbeck, Alois Pichler, Daniel Bruder, Stefan Schweter. 2016. *New (Re)Search Possibilities for Wittgenstein's Nachlass II: Advanced Search, Navigation and Feedback with the FinderApp WiTTFind*, in: Contributions of the Austrian Ludwig Wittgenstein Society, 90–93, Kirchb. A. W.

Claus Huitfeldt. 2006. *Philosophy Case Study*, in: Electronic Textual Editing, Modern Language Association of America, 181-196.

Antony Kenny. 1976. *From the Big Typescript to the Philosophical Grammar*, in: J. Hintikka (ed.), Essays on Wittgenstein in Honour of G. H. Von Wright, Acta Philosophica Fennica, 28, 41–53.

Alois Pichler and Amelie Zöllner-Weber. 2013. *Sharing and debating Wittgenstein by using an ontology*, Literary and Linguistic Computing, 28 (4), 700–707.

Alois Pichler and Tone Merete Bruvik. 2014. *Digital Critical Editing: Separating Encoding from Presentation*, in: D. Apollon, C. Bélisle, Ph. Régnier (eds.), Digital Critical Editions, 179–199, Urbana Champaign.

Alois Pichler. 2019. *A brief update on editions offered by the Wittgenstein Archives at the University of Bergen and licences for their use (as of June 2018)*, in: Wittgenstein-Studien, 10(1), 139–146.

Elena Pierazzo. 2009. *Digital genetic editions: the encoding of time in manuscript transcription*, in: M. Deegan, K. Sutherland (eds.), Text Editing, Print and the Digital World, 169–186, Famham.

Ines Röhrer. 2019. *Lexikon, Syntax und Semantik – computerlinguistische Untersuchungen zum Nachlass Ludwig Wittgensteins*, Master's thesis at LMU München, Munich.

Peter Suber. 2003. *Removing the Barriers to Research: An Introduction to Open Access for Librarians*, in: College & Research Libraries News, 64, 92–94, 113 [unabridged online version at http://legacy.earlham.edu/~peters/writing/acrl.htm].

Sabine Ullrich. 2019. *Boosting Performance of a Similarity Detection System using State of the Art Clustering Algorithms*, Master's thesis at LMU München, Munich.

Ludwig Wittgenstein. 2000. *Wittgenstein's Nachlass: The Bergen Electronic Edition*, ed. by the Wittgenstein Archives at the University of Bergen under the direction of Claus Huitfeldt, Oxford.

Ludwig Wittgenstein. 2015–. *Wittgenstein Source Bergen Nachlass Edition*, ed. by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler, in: Wittgenstein Source (2009–) [wittgensteinsource.org], Bergen.

Ludwig Wittgenstein. 2016–. Interactive Dynamic Presentation (IDP) of Ludwig Wittgenstein's philosophical Nachlass [http://wittgensteinonline.no/], ed. by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler, Bergen.

G.H. von Wright. 1969. *The Wittgenstein papers*, The Philosophical Review 78(4), 483–503.

[5] Google Analytics lists for `http://wittgensteinonline.no/` more than 4 800 and for `http://wab.uib.no/sfb/` more than 1 400 users since 2017. I would like to thank my Bergen colleagues Nivedita Gangopadhyay, Øyvind Gjesdal and Hemed Al Ruwehy for comments on a draft of this paper.

Using the FLAT repository: Two years in

Paul Trilsbeek
The Language Archive
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
Paul.Trilsbeek@mpi.nl

Abstract

In the beginning of 2018, The Language Archive at the Max Planck Institute for Psycholinguistics (TLA) migrated to a new CLARIN-compatible repository solution that was based on the open source Islandora/Fedora framework. This new solution - labeled FLAT - was developed together with the Meertens Institute and contains a customized ingest pipeline as well as a deposit front-end that is easy to use such that researchers can deposit their own collections. In the beginning of this year, some major new functionality was added to the repository for browsing and visualizing archived materials. After two years of using FLAT, in this paper we will take stock and describe what our experiences have been. What worked well? What could be improved? What turned out to be a serious shortcoming?

1 Introduction

After more than a decade of using an in-house developed archiving framework, The Language Archive decided in 2014 to start the development of a new solution that would be largely based on an existing open source repository solution (cf. Windhouwer et al. 2016, Trilsbeek and Windhouwer 2016). This was mainly done in order to reduce development and maintenance costs and to provide a better user experience. Several open source alternatives were compared and ultimately the Fedora Commons[1] / Islandora[2] combination was chosen as the option that fitted best with the requirements. Islandora is basically a set of modules for the Drupal[3] content management system and a piece of middleware to communicate with the Fedora Commons repository back-end[4]. The development was jointly undertaken with the Meertens Institute, who had similar needs. It focused on adapting the solution such that it could handle CMDI[5] metadata, and on the development of a customized ingest back- and front-end in order to have the desired level of control and ease of use. In the beginning of 2018, the new solution – labelled FLAT – was ready for production use, hence TLA was migrated to it. At that time, several performance bottlenecks that had been identified in the Fedora/Islandora framework had already been addressed. These include:

[1] https://duraspace.org/fedora/
[2] https://islandora.ca
[3] https://www.drupal.org
[4] FLAT uses the Drupal 7 version of Islandora that works with Fedora Commons 3.8.1.
[5] https://www.clarin.eu/content/component-metadata

- **Slow Resource Index.** Fedora Commons includes a so-called Resource Index, which is a combination of a relational database and a triplestore (RDF[6] database). The Resource Index can be queried when certain information about stored objects or relations between stored objects in Fedora is required. Fedora 3.8.1, which is the last version of Fedora Commons before the complete rewrite, includes the Mulgara[7] triplestore implementation. Compared to more recent triplestore implementations such as Blazegraph[8] or Apache Jena Fuseki[9], Mulgara is rather slow. Fortunately, members of the Islandora community have written alternative triplestore adapters such that Mulgara can be swapped out with "Sesame Sail"-compliant[10] triplestores such as Blazegraph or Fuseki. TLA uses Blazegraph. Search and browse performance in Islandora can be sped up significantly by deploying an Apache Solr[11] search index. Apache Solr is a very fast and highly optimized search platform that is widely used for website searches and faceted browsing. It is supported by a standard module that is part of Islandora, and many other Islandora modules can make use of the Solr index in case it is present. TLA uses Solr queries instead of queries to the Fedora Resource Index whenever possible. In addition to indexing the standard Fedora object properties in Solr, TLA also indexes all fields of all CMDI metadata files it has in the archive and makes these available in a full-text "simple search" form, as well as a selected number of fields in the form of "facets", similar to the CLARIN VLO[12]. Islandora also includes an "advanced search" option for Solr, which enables users to search in specific fields or combination of fields. We are currently looking into adding this feature to the repository.
- **Slow ingest for large objects or large batches.** An often-heard complaint about Fedora 3 is the slow performance when ingesting large objects or large numbers of objects. In the FLAT setup, this was mitigated by using the "Doorkeeper"[13] ingest workflow engine and by storing the actual archived files as "externally referenced" datastreams in Fedora. This meant that the files themselves do not actually have to pass through Fedora's API, instead pointers to the files on disk are maintained. Another advantage of this approach is that a more human friendly file system organization can be maintained, instead of the Akubra[14] low-level storage that Fedora uses by default, which uses directory names that are derived from hashes of the object identifiers. Other than the difference in low-level storage, externally referenced datastreams have the same properties and can be used in the same way as internally managed datastreams.

With those bottlenecks out of the way, we found overall performance to be good enough for production use on a large repository such as TLA with more than a million archived objects.

2 The Good

After more than two years of using the FLAT solution, TLA is still overall very pleased with the general performance, stability and usability. Searching and browsing and accessing data is generally fast and users report that they find the system easy to use, in particular the web-based deposit facility in comparison to the previous LAMUS deposit system (Broeder et al. 2006). Depositors appreciate the web-based metadata editing forms, even though these still lack any kind of batch editing options.

What is particularly convenient about the highly modular approach taken by the Islandora setup is that it is relatively easy to add or swap out certain functionality by developing new Drupal modules. Over the past two years, TLA has developed several modules for adding functionality to the FLAT setup. These were added to the production setup in the beginning of 2020 and include:

[6] https://www.w3.org/RDF/
[7] http://mulgara.org
[8] https://blazegraph.com
[9] https://jena.apache.org/documentation/fuseki2/
[10] http://archive.rdf4j.org/system/ch05.html
[11] https://lucene.apache.org/solr/
[12] https://vlo.clarin.eu
[13] https://github.com/TLA-FLAT/DoorKeeper
[14] https://wiki.lyrasis.org/display/AKUBRA

- A module for viewing ELAN[15]-annotated media in the browser
- A module for displaying access conditions with colour coded labels
- A module for displaying an image gallery for JPEG and TIFF images, using the OpenSeadragon[16] image viewer and Cantaloupe IIIF image server[17]
- A module for adding collections or objects to a "basket" and downloading them in one go as a ZIP file

The fact that the repository front-end is a widely used web content management system also makes it very easy to add basic web content alongside the repository content, e.g. for adding manuals or other types of information pages. Drupal also has a vast array of community-contributed extension modules that can be used to add functionality to the front-end.

The clear separation between front- and back-end made it possible to create two separate "portals" for different types of data in the repository, by creating a second instance of Drupal that connects to the same Fedora Commons repository back-end. One of these portals focuses on the actual language data collections, whereas the other one focuses on the other language-related data that are collected within the Max Planck Institute for Psycholinguistics, which includes for example neurobiological and genetics data.

3 The Bad

One initial shortcoming of the FLAT solution that became apparent after some period of use is that the display of so-called "compound objects" with many children is extremely slow. Compound objects are used in the Islandora setup to bundle archived objects that belong together, e.g. the different pages of a book, or a media file and its transcript. TLA uses compound objects for all bundles of resources, i.e. for all metadata records that describe resources, rather than the collection-level objects/metadata. It turned out that if a bundle has more than several hundred resources, the generation of the display for those children is very slow, as it gets a part of the required information from the Fedora Resource Index rather than the much faster Solr index. TLA has about 250 resource bundles with more than 300 resources (out of more than 110.000 metadata objects in total), for which this slow display was a problem. We expect that this issue was also the cause of the occasional slowdown of the system when it was visited too frequently by the link checking module of the CLARIN CMDI Curation Module from the Austrian Academy of Sciences, but this remains to be verified.

Fortunately, we were able to develop an alternative for the standard Islandora Compound Object viewer that uses the Solr index exclusively and speeds up the generation of the display for larger resources bundles enormously. This alternative module was installed in the TLA production environment in the middle of April, 2020.

Another issue that became apparent after some period of use is that the OAI-PMH[18] provider that is part of the standard Islandora set of modules returns server errors from time to time. The exact cause of this problem is not known yet. How serious this problem is in practice depends a bit on what an OAI harvester does in the case such an error occurs. Unfortunately, at the moment, the CLARIN VLO harvester switches to harvesting OLAC[19] records instead of CMDI records after 5 failed attempts within a period of a couple of seconds. Some more detailed logging and monitoring is needed to get to the root of this issue.

[15] https://archive.mpi.nl/tla/elan
[16] https://openseadragon.github.io
[17] https://cantaloupe-project.github.io
[18] https://www.openarchives.org/pmh/
[19] http://www.language-archives.org/OLAC/metadata.html

4 The (Somewhat) Ugly

Fedora 3 uses a system for authorisation records that makes use of the XACML[20] standard. This rather unwieldy XML standard allows for a very fine grained specification of access and management permissions on Fedora objects or even on datastreams within objects. As such, the system is more than adequate for representing the rather fine grained access polies that The Language Archive requires for its collections. In order to use the system in a way that fits TLA's needs however, each Fedora object needs to have a separate XACML policy applied to it. Applying or modifying access policies entails ingesting or modifying these XACML XML files in Fedora. For large collections with thousands of files, this turns out to be a very time consuming and computationally intensive task. This in itself is bothersome, but the main issue is that the Drupal batch process that is supposed to complete this task fails unpredictably during large modification jobs. This could be after several hours, when 80% of the job is completed. A workaround is to choose smaller parts of collections in order to have smaller batch jobs, or to use scripts that interact directly with the Fedora API, rather than using the Drupal interface and batch functionality.

To overcome this problem, we have developed a Drupal module that uses a Gearman[21] job server in the background for completing the access policy modification jobs. Even though this doesn't speed up the process very much, it at least completes it reliably. A positive outlook is that the new Fedora (4/5/6) implementation uses a more elegant approach for assigning access policies. It uses WebACL[22] policies that can be assigned anywhere in the collection hierarchy. If an object itself doesn't have a policy, Fedora walks up the hierarchy until it encounters one. This means that for a typical collection in TLA, only a handful of policies would be needed, instead of hundreds or thousands for each individual object in a collection. Modifications of those policies should then also be proportionately faster.

5 Conclusions

After more than two years of using the FLAT repository solution, the experience has been predominantly positive. Performance, stability and usability overall have been very good and the modular setup made it possible to easily add functionality. The separation between Drupal front-end and Fedora back-end made it possible to create two separate front-end "portals" that focus on different types of data in the repository. Several issues became apparent though with the actual intensive use of the solution. The display of collection objects with more than a few hundred resource was unacceptably slow and needed a customised solution that circumvented the Fedora Resource Index in favour of the Solr index. The OAI-PMH provider errors need further investigation to get to the cause of the problem. Finally, the batch modification of access policies on large collections turned out to be very slow and unreliable when initiated from the Drupal interface. Workarounds are available and a new module will at least ensure that the batch jobs complete without errors.

Drupal 7, which forms a crucial part of the current FLAT implementation and is the interface that is exposed to the world, will reach its "End of Life" in November 2022. After this, it will no longer receive security updates. Continuing to use it after this date is a considerable security risk, given that Drupal is a fairly popular web content management platform. We therefore plan to migrate to the latest Islandora and Fedora versions before that date. Islandora 8 and Fedora 4/5/6 are complete rewrites compared to Islandora 7 and Fedora 3. Migration tools are available, but we still expect this to be a fairly involved process, in particular to adapt our custom components to the new APIs and the new Drupal version.

[20] http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html
[21] http://gearman.org
[22] https://www.w3.org/wiki/WebAccessControl

Reference

Broeder, D., Claus, A., Offenga, F., Skiba, R., Trilsbeek, P., & Wittenburg, P. (2006). LAMUS: The Language Archive Management and Upload System. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006) (pp. 2291-2294).

Trilsbeek, P., & Windhouwer, M. (2016). FLAT: A CLARIN-compatible repository solution based on Fedora Commons. In Proceedings of the CLARIN Annual Conference 2016. Clarin ERIC.

Windhouwer, M., Kemps-Snijders, M., Trilsbeek, P., Moreira, A., Van der Veen, B., Silva, G., & Von Rhein, D. (2016). FLAT: Constructing a CLARIN Compatible Home for Language Resources. In K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, & A. Moreno (Eds.), Proccedings of LREC 2016: 10th International Conference on Language Resources and Evalution (pp. 2478-2483). Paris: European Language Resources Association (ELRA).

Building a home for Italian audio archives

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

Niccolò Pretto
Institute of Computational Linguistics, "A. Zampolli"
CNR, Pisa, Italy
niccolo.pretto@ilc.cnr.it

Monica Monachini
Institute of Computational Linguistics, "A. Zampolli"
CNR, Pisa, Italy
monica.monachini@ilc.cnr.it

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

Silvia Bianchi
DSFUCI, Siena University
Arezzo, Italy
bianchi.silvia83@gmail.com

Pierangelo Bonazzoli
Unione dei Comuni Montani del Cosentino, Italy
pierangelobonazzoli@casentino.toscana.it

Abstract

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

1 Introduction

Audio and audiovisual archives are scattered all over the Italian peninsula, from researchers' private houses, to universities and research centres, from cultural institutions (e.g., Istituti per la Resistenza), to State institutions, such as State Archives and Libraries. The pilot survey made by Galatà and Calamai in 2018 has emphasised the status of precariousness, instability, and insecurity that affects audio and audiovisual archives available at different communities of Italian researchers. Almost half of the resources listed in survey (49.6%) were barely accessible. Only 9.2% of the resources was accessible and available, 4.6% was partially accessible, 35.1% was available upon request, 1.5% is available upon request and only for selected parts. As for the resources which were declared to be accessible, the access policies were as follows: only 9.2% of these resources was freely accessible online (with no authentication); 7.6% was accessible online via authentication; 29% was accessible onsite (i.e. where the resources are physically stored). As for the long-term maintenance and preservation the answer receiving the highest number of responses was nobody (43%), followed by reference Institutes, such as Associations, Foundations, libraries and their archives (17%), reference Universities (16%), the owners/individuals themselves (15%)[1]. Several research projects in recent years aimed to disseminate audio and audiovisual archives, which are collected over the years by both researches and amateur fieldworkers: some examples are, among others, *Grammo-foni. Le soffitte della voce*, also referred as Gra.fo (Calamai and Biliotti, 2017), *Voci, parole e testi della Campania*[2], *I granai della memoria*[3], *Circolo Gianni Bosio Audio Archives*[4]. Nevertheless, fragmentation and lack of common and shared standards are often the common features of certain initiatives, whose duration over time crucially appears to be dependent on the duration of external funding, if any. Moreover, a researcher working with audio archives is not necessarily competent in long-term preservation of audio data and data management. Eventually, not all the research projects

[1] Further details available in (Galatà and Calamai, 2019)

[2] www.archivicampani.unina.it/archivi_campani_dev Last visited August 20th, 2020

[3] www.granaidellamemoria.it Last visited August 20th, 2020

[4] www.circologiannibosio.it/archivio.php Last visited August 20th, 2020

dealing with audio archives receive financing for all the different professional profiles involved in their preservation, managing and valorisation.

Given this picture, it appears urgent to provide an infrastructure offering: 1) a long-term preservation service for audio archives, 2) a shared set of metadata compliant with the main international standards and FAIR principles, and 3) an access interface which takes into account the peculiarities of the audio modality and which is able to support researchers in different disciplines. This paper presents how the Archivio Vi.Vo. project tackles these problems, illustrating the overall adopted methodology and the developed infrastructure.

2 The Archivio Vi.Vo. project

2.1 The project

In 2019, Regione Toscana decided to support the project Archivio Vi.Vo., which aims to catalogue and disseminate oral archives. Partners involved are: Siena University (Silvia Calamai), CNR-ILC & CLARIN-IT (Monica Monachini), Soprintendenza Archivistica e Bibliografica della Toscana (Maria Francesca Stamuli) and Unione dei Comuni del Casentino (Pierangelo Bonazzoli). In order to reach the above ambitious objectives, Archivio Vi.Vo. concentrates most of the efforts on the design and development of an architecture, hosted by CLARIN-IT, the Italian node of the CLARIN research infrastructure, that could be used by several other projects concerning audio archives. A crucial step towards this main aim concerns the definition of the metadata set(s) used to describe the data. The set has to be compliant with international archival standards, such as with ISAD(G) and ISAAR, as well as several others derived from different disciplines (cf. below Section 3.1).

2.2 The case study

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

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

In several cases, the original carriers were devoid of all the contextual information (place and date of recordings, speakers involved in the recordings). In other cases, the open-reel tapes were recorded at different speeds and using different track head configurations, thus making rather complex the digitisation process and the creation of access copies. From this respect, Caterina Bueno's audio archive represents

an extreme test case where different levels of complexity call into question different professional profiles and skills.

3 Building the home for audio archives

3.1 Data and metadata

In order to preserve and provide access to analogue audio recordings (e.g., the compact cassette or the open reels), it is essential to digitise them. The result of the digitisation process is the digital *preservation copy*, which is composed by the audio content as well as several other information about the carrier (such as the photo of the carrier itself, or its box)[5]. As the name suggests, the preservation copy is the "means" for safeguarding the content of the audio documents and it can be considered as the new digital master for long-term preservation. Another important concept to be defined is the *archival unit*. In audio and audiovisual archives, it is defined as a set of data and documents pertaining to the very same communicative event, per unit of time and place. Archival unit is the outcome of a meticulous process involving listening, analysis and comparison. Sometimes audio content needs to be re-organised. For example, an archival unit could be composed by content that is stored in several physical carriers (and, therefore, in several preservation copies), or vice versa, several archival units could be stored in the same physical carrier[6]. Given the absence of a one-to-one relationship between the physical carrier (i.e., compact-cassettes, open-reels) and the archival unit, the preservation copies are kept separately from the archival units (Mulè, 2003; Calamai et al., 2014; Stamuli, 2020).

This approach leads to a very complex set of metadata, articulated along three different layers: (i) metadata for the description of the preservation copy, (ii) metadata for the description and managing of oral sources as items of an (audio) archive (archival unit), and (iii) metadata expressing the relationship between the preservation copy metadata and the digital archive metadata.

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

3.2 From preservation copies to archival units

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

Sometimes the audio recording is not easily accessible, due to, e.g., different speeds, configurations or digitisation errors (Pretto et al., 2020). In these cases, if necessary, researchers or audio technicians recur to the concept of clip[7] in order to separate parts with different speeds, channels with different recordings or recordings in different directions. In Archivio Vi.Vo. a clip is defined as a duplicate of an audio segment extracted from a preservation copy. One or more clips can be extracted from a preservation copy. In some cases, the clips are the result of a restoration operation, necessary for the use of the sound content. The process of creating (and restoring) the clips must in no way modify the preservation copy.

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

[6] Several other kinds of transformation could be performed, but their description goes beyond the scope of the paper

[7] The concept of clip is commonly used to indicate data of either video or audio that has been clipped out (copied) from a larger environment such as a reel or a video tape

The resulting clips will be correctly accessible and allow the researcher/cataloguer to listen, analyse and describe their contents. In case some parts of the very same clip belong to different events (and, therefore, to different archival units, see Section 3.1), they will be segmented accordingly and new sub-clips will be created (archival unit clips). In some cases, some archival unit clips, derived from different preservation copies, would be part of a same event, therefore, they will also be part of the same archival unit. As soon as all the archival unit clips will be ordered, and all the missing metadata required by ISAD(G) will be added, the archival unit will be created and available through the access interface.

3.3 The infrastructure

As for the infrastructure, ILC4CLARIN and the CLARIN-IT national data centre (Monachini and Frontini, 2016), will implement new experimental approaches to preservation, management and access to audio data and metadata. The experimental activity aims to adopt the model and the high-performance computing and archiving services of the new GARR network infrastructure, built along the Cloud paradigm[8]. The project will also exploit the federated identity service of the CLARIN infrastructure, in order to manage users' access. A robust system for managing authentication is essential for audio and audiovisual archives because of the frequent privacy, ownership, and copyright issues concerning their content (Kelli et al., 2019; Kelli et al., 2020). Several classes of users are considered each of them with different access grants.
The infrastructure consists of two different parts. The first one provides a data and metadata entry interface for archivists, archive owners or, in general, researchers who want to preserve the legacy. The system is highly complex: it must be able to manage several international standards and several kinds of specific functionalities. Considering the complexity of the project, the infrastructure would difficulty be developed from scratch. Therefore, as a first step, ten archival software were evaluated on the base of several features and technologies (standards, programming languages, frameworks, DBMS, license, etc.). The selected software was the open-source software xDams[9]. Three main characteristics influenced the adoption of the software: (i) the completeness of its standards coverage, (ii) its extensible no-sql database as well as (iii) the open-source license. The second part of the infrastructure consists of an access interface able to support researchers of different disciplines in discovering and studying audio or audiovisual documents. In order to study the interaction with the software, two mockups were developed for studying and testing the interfaces for inserting and cataloguing the digitised documents (Figure 1a) and for accessing their content (Figure 1b), respectively. The two mockups have been developed with the frameworks Vue.js and Bootstrap, respectively, Web Audio API, as well as Peak.js and Audiowaveform, two libraries developed by BBC[10].

4 Final remarks

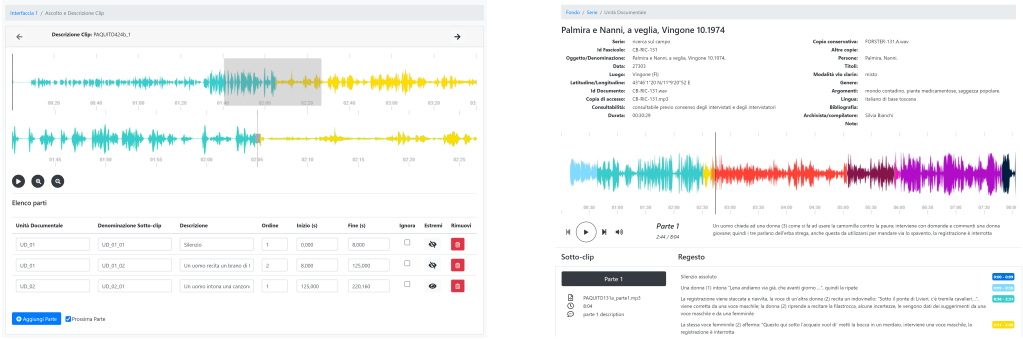
Archivio Vi.Vo. constitutes a pilot case study within CLARIN-IT to experiment with methods for long-term preservation and secure access to oral source and offer targeted services for both specialists and the general public interested in these data. Archivio Vi.Vo. aims not only to develop an infrastructure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

References

Federica Bressan and Sergio Canazza. 2013. A systemic approach to the preservation of audio documents: Methodology and software tools. *Journal of Electrical and Computer Engineering*, 2013:21 pages.

Silvia Calamai and Francesca Biliotti. 2017. The gra.fo project: from collection to dissemination. *Umanistica Digitale*.

[8] `cloud.garr.it` Last visited August 20th, 2020
[9] `www.xdams.org` Last visited August 20th, 2020
[10] `github.com/bbc/peaks.js` and `github.com/bbc/audiowaveform` Last visited August 20th, 2020



(a) (b)

Figure 1: (a) section of the interface for creating archival units from preservation copies: the clips derived from the duplicate of preservation copy are segmented, described, ordered and assigned to an archival unit; (b) the interface for accessing the archival units contents

Silvia Calamai, Francesca Biliotti, and Pier Marco Bertinetto. 2014. Fuzzy archives. what kind of an object is the documental unit of oral archives? In Marinos Ioannides, Nadia Magnenat-Thalmann, Eleanor Fink, Roko Žarnić, Alex-Yianing Yen, and Ewald Quak, editors, *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*, pages 777–785, Cham. Springer International Publishing.

Vincenzo Galatà and Silvia Calamai. 2019. Looking for hidden speech archives in italian institutions. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, number 159, pages 46–55. Linköping University Electronic Press.

IASA Technical Committee. 2009. *Guidelines in the Production and Preservation of Digital Audio Objects: standards, recommended practices, and strategies. IASA TC-04*. International Association of Sound and Audiovisual Archives, 2nd edition.

Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labropoulou, Maria Gavriilidou, Pavel Stranák, et al. 2019. Processing personal data without the consent of the data subject for the development and use of language resources. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, Italy, 8-10 October 2018*. Linköping University Electronic Press.

Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Straňák, et al. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. *Selected Papers from the CLARIN Annual Conference 2019, Leipzig, Germany, 30 September, 2019*.

Monica Monachini and Francesca Frontini. 2016. CLARIN, l'infrastruttura europea delle risorse linguistiche per le scienze umane e sociali e il suo network italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics*, 2(2):11–30.

Antonella Mulè. 2003. Le fonti orali in archivio. un approccio archivistico alle fonti orali. *Archivi per la storia*, 16(1):111–129.

Niccolò Pretto, Alessandro Russo, Federica Bressan, Valentina Burini, Antonio Rodà, and Sergio Canazza. 2020. Active preservation of analogue audio documents: A summary of the last seven years of digitization at csc. In *Proceedings of the 17th Sound and Music Computing Conference, SMC20, Torino*, Torino.

Maria Francesca Stamuli. 2020. Fonti orali, documenti e archivi: riflessioni e proposte per la nascita di un 'archivio vivo'. In Duccio Piccardi, Fabio Ardolino, and Silvia Calamai, editors, *Studi AISV 6*.

Digitizing University Libraries – Evolving From Full Text Providers to CLARIN Contact Points on Campuses

Manfred Nölte
State and University Library
Bremen, Germany
noelte@suub.uni-bremen.de

Martin Mehlberg
State and University Library
Bremen, Germany
martin.mehlberg@suub.uni-bremen.de

Abstract

The beginnings, currently emerging demands and activities as well as future options of the relation of the State and University Library Bremen (SuUB) to CLARIN will be described in this paper. Section 1 presents a solution for supplying digital humanists with tools and services suited best for their research. With the SuUB as a digitizing academic library this relation is with respect to full text transfers to CLARIN (Geyken et al., 2018; Nölte and Blenkle, 2019). Further connections could consist in providing advice and training for researchers of the Digital Humanities as potential CLARIN users (see section 2) and a discussion about future structural options on the level of research infrastructures. In section 3 we suggest a collaboration between digitizing libraries to jointly agree upon standards of quality, file formats, interfaces and web services. We discuss the foundation of local CLARIN contact points to pass on scholars and researchers to the respective contact or service of CLARIN. The relevance to the CLARIN activities, resources, tools or services is described at the end of each respective section.

1 Digitizing University Libraries as Full Text Providers for CLARIN

The State and University Library Bremen is one of many libraries dedicated to the digitization of its historical collections. Digitization and especially the generation of full text is an important instrument for improving the accessibility of valuable information contained in fragile historical documents. It facilitates academic research and teaching and is indispensable to the digital humanities.

Usually, digitizing university libraries produce digital images, metadata for cataloguing and web-navigation purposes, and optical character recognition (OCR) full text for searching. This information is made available through the library's web portal for digital collections. However, digital humanists need high-quality full texts enriched with metadata in the appropriate format to analyze them with powerful software tools. The SuUB has actively transferred full texts created by digitization projects (funded by the German Research Foundation, DFG) to the research infrastructure CLARIN. We would like to outline our approach adopted so far, the results and the dissemination achieved within the scientific community. Later on in section 3 we discuss the underlying structure and concept and how we might intensify actions like this.

The historical journal *Die Grenzboten* was the first full text transferred from the SuUB to CLARIN (Geyken et al., 2018). *Die Grenzboten* is a long running serial publication (1841–1922). It can be classified as a literary journal that also covered politics and arts. It was founded by Ignaz Kuranda (1811-1884) in 1841 in Brussels and later on published in Leipzig and Berlin. We demonstrate that good OCR quality and a page-wise structuring are prerequisites for the creation of a high-quality Text Encoding Initiative (TEI) version of a full text. The TEI version was created in cooperation with the Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) (Nölte et al., 2016).

We digitized more than 185,000 single pages in 270 volumes. Almost 33,000 articles were digitized via optical character recognition (OCR) and the titles of the articles were manually captured. The resulting OCR full text was processed by the OCR software ABBYY Finereader 9 and consists of

approximately 500 million characters and 65 million tokens. As a second aspect of text quality we enhanced the level of document structure according to an agreed standard format together with our partners, the Deutsches Textarchiv (DTA) (Haaf Geyken and Wiegand, 2014/15). Using this structure information we converted the OCR output format to an interoperable TEI-format. The metadata of the 33,000 articles also contain information about the publication dates, so that it is possible to analyse the full texts over time.

We were active to disseminate our digitized historical journal, as has been shown. It has been used as textual material by computer linguists, digital humanists and philologists as well as it has been part of academic teaching within universities like Ghent, Würzburg and Göttingen. Together with ground truth full text data it has been used by OCR post correction system providers and big projects like OCR-D (Neudecker et al., 2019). The journal *Die Grenzboten* was subject to diachronic collocation analyses[1] as well as to analyses like topic modeling (Jannidis, 2016).

Later on the SuUB and the University Library Frankfurt digitized over 1000 book titles with about 245,000 pages within the project "Digitale Sammlung Deutscher Kolonialismus" (digital collection of texts from the period of German colonialism). This project also generated full texts that have been transferred to CLARIN. Both full text transfers are showcases for a collaboration/teamwork between a digitizing library and CLARIN. The question now is: What has to be done to intensify the transfer of huge amounts of digitized full texts in a reasonable and cost-effective routine manner? Below, in section 3 we will come back to this question in more detail.

Relevance to the CLARIN activities: As described above, the SuUB has actively transferred full texts to CLARIN in an early phase directly after the digitization process. Doing so, we helped to increase the amount of **language resources provided by CLARIN** and together we generated a bigger perception of this full text within the scientific community. **CLARINs "ingest service"** and the possibility to host the full text in the **CLARIN repositories** also clearly helped to increase the dissemination of our full texts within the scientific community.

Digitization activities have the potential to create very huge amounts of digitized full texts. This will stimulate **inter- and cross-disciplinary research**. This first collaboration between CLARIN and the SuUB Bremen can be seen as a **showcase scenario** of how content providing libraries and CLARIN can mutually benefit from this kind of digitization projects.

2 Counselling and Training Activities of the SuUB Bremen

Academic libraries are close to scholars and researchers not only in terms of physical closeness but also in terms of subject proximity (providing information and services). Libraries have been and are the contact point for diverse questions and demands regarding literature supply and also constitute a so called "learning space" for students and scientists.

Libraries of course know best "their own material", i.e. texts digitized by themselves. They know the subject, the context and the quality of the digitized material, which expresses itself in the quality of the originals, pixel images, error rates of the full texts as well as the scope and the quality of the metadata. Further technical issues are the interfaces to get access to the data and the formats delivered by the respective systems. This enables the libraries to help with quality issues or options to use web services like IIIF. For example, the quality of the section titles within the full text of *Die Grenzboten* is poor, due to a usage of a special type, whereas the same information contained in the METS-xml files has a perfect quality, because it was captured manually.

In the past the SuUB has gathered some experience with the personal counselling of scholars from across the humanities disciplines, including linguists and political scientists. In general, the questions were of a technical nature (relating to e.g. formats or system interfaces), but sometimes also questions of a more theoretical manner were discussed (e.g. kinds of quantitative analyses like topic modeling, diachronic collocations, etc.).

The more technical counselling aims to obtain interoperable full texts that meet the requirements of the software tools (like regular expression search, part-of-speech tagging, named-entity recognition or

[1] Jurish, Bryan, M. Nieländer, and T. Werneke. 2017. "DiaCollo and *die Grenzboten*." Talk presented at the conference GenealogJurish, B., M. Nieländer, and T. Werneke. "DiaCollo and die Grenzboten." Talk presented at the conference Genealogies of Knowledge I: Translating Political and Scientific Thought across Time and Space, University of Manchester, 7th-9th December, 2017ies of Knowledge I: Translating Political and Scientific Thought across Time and Space, University of Manchester, 7th-9th December. (pdf:abstract, pdf:slides)

topic modeling). Especially with structured full texts (TEI or in general all sorts of xml) format issues have to be considered. Some quantitative tools, like mallet (topic modeling) only need plain text. But the pre-processing or the whole tool chain (e.g. including a graphical presentation for the analysis findings) nearly always requires the above mentioned features: structured pages (i.e. semantically tagged full text) and metadata (year of publication, authors, etc.).

Relevance to the CLARIN activities: As shown digitizing libraries are in a good position to start **Researcher training activities** with respect to their full text resources. Furthermore they can help accessing **web services or metadata offered by the digital collections software systems**, like IIIF and OAI-PMH.

Actively supporting the above mentioned full text transfers and the mentioned counselling activities will result in considerably better outcomes in all fields of automated and computer-aided research across disciplines working with digitized material. It will enable the employment of quantitative methods and approaches such as authorship attribution studies, clustering techniques (i.e. for literary genre analysis), topic modeling etc.

3 Prospects for Future Collaboration Between CLARIN and Academic Libraries

As shown above digitizing libraries already play a role in the context of CLARIN and the group of CLARIN users. The next step should be to intensify the collaboration between CLARIN and academic digitizing libraries in order to harmonize the transfer of digital textual material, to jointly agree upon common activities or even to establish CLARIN contact points on the campuses of the universities. And the most appropriate place is a library as we will demonstrate below.

Having done the above mentioned full text transfers a few times we list some criteria that might be in need of a more precise specification together with potential requirements. Here we give only short explanations, see (Nölte and Blenkle, 2019) for more examples and details.

- *Full text quality*: For example a maximum error rate of characters or for the structuring of the text. These criteria vary for different centuries or decades or different software tools or scientific approaches.
- *File formats and metadata*: Transferring plain text is not an option, nor is the output of OCR engines (such as ABBYY-XML) as well. There has to be a decision for ALTO, TEI or other file formats, possibly together with ‘annotation guidelines’. (Haaf et al., 2014/15)
- *Persistent back links*: Within the full text there should be back links (page-wise or at least by sections) to the scanned images in the digital collections of the respective library or archive. If possible, these back links should be persistent at a URN granular level.15 Researchers appreciate having the possibility to check the original image quality or to have access to supplemental material such as graphics, images, advertisements, or vignettes.
- *Line breaks*: There should be a guideline for whether to transcribe line breaks as is. We have cooperated with partners with varying opinions on this question. Some institutions wanted line breaks as is, whilst the transcription for Wikisource had to be without wrapped words.
- *Strictness of character transcription*: Within historical full texts the spelling, of course, should be transcribed as is; for instance, ‘Säugethiere’ with ‘th’ and ‘Entwickelung’ instead of the modern form ‘Entwicklung’. The same should apply to the transcription of single historical characters using UTF8 codes, like ligatures or special historical glyphs.

Ideally, there should be a documentation listing all the above-mentioned information: the level of the ‘full text and metadata’ quality, whether there are further file formats available, the availability of back links, and the status of line breaks and character transcriptions. If this ‘full text metadata’ would be realized with computer readable XML formats, pre-processing scripts might automatically decide what pre-processing remains to be done, and what analysis tools or scientific approaches might be applicable. Licensing and intellectual property would be a further major issue to address.

Jointly discussing and agreeing upon criteria and requirements like this will lead to a best practice approach for future transfers of full texts to CLARIN. Another future activity could consist in the establishment of CLARIN contact points for scholars and scientists at academic libraries. These already have a proficiency in counselling and offering services in the respective domains. They function as well as learning spaces, aiming at creating the best atmosphere for the exchange of knowledge. Other outstanding advantages of libraries are: Libraries constitute sustainable structures in

the scientific world, they are research infrastructures themselves and they are local. Every university has a library and universities are places where students get scientists which might get passed on as potential users of CLARIN.

Currently there are a lot of activities to establish event formats at libraries as scholar or researcher training activities with names like "digital lab", "hands-on lab", "GLAM lab", "innovation lab", "data lab", "HackyHour", "digital learning lab", "digital humanities lab", "library lab", "scholarly makerspaces" or more combinations of these words. CLARIN might play a role to harmonize this multitude of activities, to combine these with CLARIN's experience, services and tools. The above mentioned start of a collaboration between libraries and CLARIN might be a good first step. Together we might even go for a fully automated full text transfer or harvesting as a long-term goal.

Relevance to the CLARIN activities: Setting up the described collaboration will **standardize** and harmonize all future full text transfers and training activities in the context of CLARIN and digitizing libraries. I.e. together we will create **best practice approaches** that provide scholars and researchers with the best possible quality and **interoperability of language resources and services**.

As shown, the SuUB together with CLARIN has a big potential to establish a **user assistance**, **help desk** or contact point. While documenting the digital collections software system with **user manuals** we might support the scholars and researchers within the domain of digital language resources. An example is information about our systems **persistent identifiers and citation mechanisms** (see above the criterion for "*Persistent back links*").

References

Geyken, Alexander, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In: Henning Lobin, Roman Schneider, Andreas Witt (Hgg.): Digitale Infrastrukturen für die germanistische Forschung (= Germanistische Sprachwissenschaft um 2020, Bd. 6). Berlin/Boston, pages 219–248. Online-Version, DOI: 10.1515/9783110538663-011.

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2014/15. 'The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources', *Journal of the Text Encoding Initiative*, no. 8, no page

Jannidis, Fotis. 2016. 'Quantitative Analyse literatischer Texte am Besipied des Topic Modeling', *Der Deutschunterricht*, 68.5, 24–35

Neudecker, Clemens, Konstantin Baierer , Maria Federbusch, Kay-Michael Würzner; Matthias Boenig, Elisa Herrmann, and Volker Hartmann. 2019. OCR-D: An end-to-end open-source OCR framework for historical documents, in: *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, Brüssel 09.05.2019, pages 53–58. Online: <https://dl.acm.org/doi/10.1145/3322905.3322917> [accessed 27 April 2020].

Nölte, Manfred and Martin Blenkle. 2019. '*Die Grenzboten* on its Way to Virtual Research Environments and Infrastructures', *Journal of European Periodical Studies*, 4.1, 19-35

Nölte, Manfred, Jan-Paul Bultmann, Maik Schünemann, and Martin Blenkle. 2016. 'Automatische Qualitäts-verbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift *Die Grenzboten*', o-bib, 3.1, 32–55 (p. 32) [accessed 27 April 2020].

"Tea for two": the Archive of the Italian Latinity of the Middle Ages meets the CLARIN infrastructure

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

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

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

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

Paolo Monella
ALIM, Università degli
Studi di Palermo, Italy
paolo.monella@unipa.it

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

Abstract

This paper presents the Archive of the Italian Latinity of the Middle Ages (ALIM) and focuses, particularly, on its structure and metadata for its integration into the ILC4CLARIN repository. Access to this archive of Latin texts produced in Italy during the Middle Ages is of great importance in providing CLARIN-IT and the CLARIN community, at large, with critically reliable texts for the use of philologists, historians of literature, historians of institutions, culture and science of the Middle Ages.

1 Introduction

The Archive of the Italian Latinity of the Middle Ages – in Italian, Archivio della Latinità Italiana del Medioevo (ALIM) – is an Italian national research project aimed to provide free online access to a large number of Latin texts produced in Italy during the Middle Ages. ALIM makes an unprecedented contribution to the studies not only of Latin, but also of culture and science at the basis of our Wester European society. The general aim of the paper is to allocate ALIM within the framework of CLARIN-IT and CLARIN at large. Section 2, shows how ALIM may contribute to fill an important gap in textual sources: query searches run on the Virtual Language Observatory for Latin-related resources demonstrate that no resource with the features and potentialities of ALIM is currently available. The technical description of the internal structure and metadata of the Archive is discussed in Section 3 while the strategy for the integration of the ALIM archive into the ILC4CLARIN repository is discussed in Section 4. Finally, ALIM's benefits for the CLARIN-IT research directions and for the CLARIN community are presented.

2 Classics resources in CLARIN

The Italian CLARIN (CLARIN-IT) consortium[1] has strong interest in the field of Digital Classics, which still suffers from shortage or restricted availability of lexical resources for historical languages.

Within the CLARIN-IT consortium, the collaboration between the Centre for Comparative Studies "I Deug-Su", Department of Philology and Literary Criticism at the University of Siena – DFCLAM[2] and the data centers mostly concerns the study of methods and services to offer online secure access to some digital archives of literary and historical texts: among them, ALIM (the Archive of the Italian Latinity of the Middle Ages), hosted by the University of Siena, http://alim.unisi.it/il-progetto/), is the largest digital library of the Italian Latinity including both literary and documentary texts.

Evidence shows that the CLARIN data centers do not offer resources such as ALIM. The faceted-search functionality of the the Virtual Language Observatory (VLO), performed combining *Latin text resource* and *Middle Ages*, returns 53 records only (124 are returned by the query *Latin* combined with the adjective *medieval*). These data are essentially images of manuscripts; no XML-TEI texts seem to be available by using these search keys. A further query with XML as free text, Latin as language, and text or corpora as resource type returns about 1300 records, mostly consisting in Treebanks or in documents coming from EUROPEANA[3].

Having ALIM in CLARIN is, thus, very important for both CLARIN and the community of users. ALIM will offer high-quality resources because of the following reasons: (i) the resources are digitized by domain experts; (ii) the large number of resources that cover a wide historical period; (iii) a strong organization dedicated to the maintenance; (iv) all the offered resources are TEI encoded.

3 ALIM: history, goals, structure

ALIM is an archive of medieval Latin texts composed in the Italian area between the $8th$ and $15th$ century. It originated as a UAN (Unione Accademica Nazionale) project in the Nineties and was later supported by the national Ministry of Education. Its original aim was twofold: to make medieval Latin literature texts openly available and to provide a textual corpus serving as a basis to create a new dictionary of medieval Latin in its Italian variety. The latter goal explains a unique feature of ALIM: it does not only include literary sources, but also collections of documentary texts. ALIM is therefore divided into two sections: "Fonti letterarie" and "Fonti documentarie". While the majority of texts are drawn from printed editions, some are new, born-digital editions[4].

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

Until 2016, ALIM was hosted by the servers of the University of Verona, Italy[5] and its texts had procedural markup, based on simple HTML markers. We shall refer to this version as "ALIM1".

In 2016, the current version of the archive ("ALIM2") was launched. The migration process involved the following tasks: (1) building a new open source software TEI XML-based digital library infrastructure and publishing it in the servers of the University of Siena[6]; (2) re-encoding text markup and metadata in TEI XML P5.

Task 1 was realised in collaboration with the external IT company Net7[7] and completed in 2016/17, when the ALIM2 website was launched. Task 2 involved a longer process, still ongoing, curated by the "équipe di codifica" (E. Ferrarini, P. Monella and R. Rosselli Del Turco) with the aim of gradually improving the level of formalization and the granularity of text markup and metadata.

[1] The composition of the Italian Consortium cf. (Nicolas et al., 2018) is available at `http://clarin-it.it/en/content/consortium`

[2] Prof. Francesco Vincenzo Stella

[3] `http://www.europeana.eu`

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

[5] `http://www.alim.dfll.univr.it/`

[6] `http://alim.unisi.it/`

[7] `https://www.netseven.it/`

In the current version of the archive, each literary text is encoded as a TEI XML P5 file with a `<TEI>` root element, while in the documentary section each TEI XML file includes a whole volume of a documentary collection[8], has a `<teiCorpus>` root element and includes each individual document in a separate `<TEI>` element. In the latter case, both `<teiCorpus>` and `<TEI>` have their own `<teiHeader>` with metadata respectively regarding the whole collection and the individual document.

In ALIM2 TEI XML files for literary texts deriving from the initial export from ALIM1 (labelled as "encoding level ALIM2_0"), much metadata was included in unstructured `<note>` elements of the TEI. Also, most texts lacked any TEI structural markup such as `<div>`. In 2017/18, literary texts were gradually upgraded to "encoding level ALIM2_1" thanks to the work of ALIM collaborator Chiara Casali on metadata integrity, and of Jan Ctibor on metadata encoding and structural markup. Jan Ctibor's activity was brought forth in the framework of a collaboration agreement between ALIM and the *Corpus Corporum*[9], the largest full-text repository for Latin (163 M words). The current policy of ALIM requires that all new texts included to the archive must be encoded at "level ALIM2_2": this includes markup of work titles, quotes, speeches, person or place names.

The archive also includes born-digital scholarly editions directly based on handwritten medieval witnesses, whose encoding level is labelled as ALIM2_3[10].

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

4 ALIM in CLARIN-IT

4.1 Structure for ALIM data into ILC4CLARIN repository

As described in Section 3, the ALIM digital library is arranged into two complementary sections: *Fonti Letterarie (Literary Sources)* and *Fonti Documentarie (Documentary Sources)*. The former is a collection of single documents (about 350), while the latter is a collection of 50 corpora that groups about 6455 texts. Since ALIM keeps these two resources separated, we decided to mirror this structure in the ILC4CLARIN repository. We created two collections, *Literary Sources* and *Documentary Sources*, under the *OPEN* community[11]. This structure is important for, at least, two reasons. The first one is that researchers accustomed to ALIM find in the ILC4CLARIN repository the same structure they are used to; the second one is connected with the VLO. In section 2, we briefly mentioned the faceted-search of the VLO. Well, one of such facets is exactly the collection (in the repository) the data come from. The ALIM data are retrieved from the VLO using either "fq=collection:ALIM+Literary+Sources&fqType=collection"or "fq=collection:ALIM+Documentary+Sources&fqType=collection"[12].

4.2 Population of the repository with ALIM data

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

[8] E.g.: *Codex diplomaticus Cavensis*, volume 1: `http://alim.unisi.it/dl/fonte_documentaria/7381`.

[9] `http://www.mlat.uzh.ch/MLS/`

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

[11] Since ILC4CLARIN uses the clarin-dspace repository, we have used the terminology community and collections. For clarity, collections are nested into communities.

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

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

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

Given the large number of items to describe in the repository, we decided to use the import functionality of the repository[13] to batch-load the items. Since this procedure is unsupervised, as far as the content of the items is concerned, we decided to manually create a prototypical item, export it, and automatically clone it. In this way, every item is syntactically correct and can be safely imported into the repository.
In details: (i) we took one document from literary sources and one from documentary works and kept them as prototypes; (ii) we carefully created a submission, mapping the elements of the `<teiHeader>` into the fields of the submission form of the repository; and (iii) once the internal workflow of metadata quality is passed, we exported the item.
The exported item is an archive which contains the following metadata files: **metadata_local.xml**, **dublin_core.xml**, and **metadata_metashare.xml**. All of them are populated with data extracted from elements of the `<teiHeader>`. The ALIM research team checked sample metadata from the CLARIN archive and verified that they correspond to those included in the TEI headers of the ALIM XML files and to the general project information pertaining to the archive. Before concluding, let's add that the official URL of the ALIM project (in our case, `http://it.alim.unisi.it/`) is contained in the **dublin_core.xml** files, while **metadata_local.xml** files contain the *demo URL* This mapping enforces our decision to describe the `<teiCorpus>` instead of describing every single document in the corpus. Literary Sources have a clear URL where the document resides: for example, the "Dialogus" by Gerius Aretinus is available at `http://it.alim.unisi.it/dl/resource/194`. By contrast, Documentary Sources point to URLs that report the whole corpus. For example, the "Codex diplomaticus Cavensis - 01" is available at `http://it.alim.unisi.it/dl/fonte_documentaria/7381`. On the web page, a JavaScript allows the user to jump to the desired documents, such as the 27^{th} document, whose internal URL is `http://it.alim.unisi.it/dl/fonte_documentaria/7381#doc_27`. Unfortunately, '#' is a reserved character[14] which separates information sent to server from client side actions, and no data transmitted as part of the URL must contain it.
The complete mapping guide, the scripts, and XSLT style sheets are available at `https://github.com/cnr-ilc/alim2clarin-dspace`.

4.3 Versioning

The ILC4CLARIN repository implements the versioning of the described items. Indeed, it is always possible to add to the repository a new item as "new version of" a previous one. The versioning of the items on the repository should be consistent with the one on the ALIM digital library. The latter allows contributors to replace the XML-TEI of a literary work or documentary collection with a new one, including changes in the text or in the metadata. The ALIM2 digital library keeps all previous XML files available in the backend, but only makes the last one (and the derivative HTML, PDF and plain text files) available to the user.
To make the versioning of the ILC4CLARIN repository coherent with ALIM's, we decide to remove the demo URL from the old version(s). In this way, the users access the last version of the document from the repository and, if they nevertheless need previous data, can contact ALIM and request for the previous data.

[13] `https://wiki.lyrasis.org/display/DSDOC5x/Importing+and+Exporting+Items+via+Simple+Archive+Format`
[14] `https://www.urlencoder.io/learn/`

5 Concluding remarks

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

ALIM complements the Latin resources in CLARIN by providing access to a large corpus of medieval literary and documentary Latin texts with granular curated metadata. On the other hand the VLO makes the resources produced and described in the ILC4CLARIN repository, including ALIM metadata, available to a wider audience in the SSH community, while the CMDI model ensures high quality metadata curation. Also, CLARIN offers ALIM the possibility to use technology and text analysis tools available at CLARIN data centers to deal with multilingual data. For example, Weblicht allows to combine web services so as to handle and exploit textual data.

References

[Alessio2003] Gian Carlo Alessio. 2003. Il progetto alim (archivio della latinità italiana del medioevo). In Francesco Santi, editor, *In Biblioteche elettriche. Letture in Internet: una risorsa per la ricerca e per la didattica*, volume 1, pages 73–81. SISMEL - Edizioni del Galluzzo.

[Buzzoni and Rosselli Del Turco2016] Marina Buzzoni and Roberto Rosselli Del Turco. 2016. Evolution or revolution? digital philology and medieval texts: History of the discipline and a survey of some italian projects. In *Mittelalterphilologien heute. Eine Standortbestimmung. Band 1: Die germanischen Philologien*, pages 265–294. Königshausen und Neumann.

[D'Angelo and Monella2019] Edoardo D'Angelo and Paolo Monella. 2019. ALIM (Archivio della Latinità Medievale dItalia). Storia, attualità, prospettive di una banca-dati di testi mediolatini. In Roberto Gamberini, Paolo Canettieri, Giovanna Santini, and Rosella Tinaburri, editors, *La Filologia Medievale. Comparatistica, critica del testo e attualità. Atti del Convegno (Viterbo, 26-28 settembre 2018)*, volume 3 of *Filologia Classica e Medievale*. L'Erma Di Bretschneider.

[Ferrarini2017] Edoardo Ferrarini. 2017. ALIM ieri e oggi. *Umanistica Digitale*, 1:7–17.

[Nicolas et al.2018] Lionel Nicolas, Alexander König, Monica Monachini, Riccardo Del Gratta, Silvia Calamai, Andrea Abel, Alessandro Enea, Francesca Biliotti, Valeria Quochi, and Francesco Vincenzo Stella. 2018. CLARIN-IT: State of Affairs, Challenges and Opportunities. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18-20 September 2017*, Linköping electronic conference proceedings (Print), pages 1–14.

Use Cases of the ISO Standard for Transcription of Spoken Language in the Project INEL

Anne Ferger
Universität Hamburg, Germany
anne.ferger@uni-hamburg.de

Daniel Jettka
Universität Hamburg, Germany
daniel.jettka@uni-hamburg.de

Abstract

This contribution addresses the benefits and challenges of using the ISO standard for transcription of spoken language ISO 24624:2016 in a long-term research project. By exploring several use cases for the standard format, which include its application in archiving, dissemination, analysis and search of linguistic data, the practicality and versatility of its usage are examined. Also various opportunities for further interdisciplinary research are highlighted.

1 Introduction

The beneficial usage of technical standards has become a theoretical commonplace in digitally oriented humanities research. There is widespread consciousness and active participation in the optimization of interoperability and long-term accessibility of created resources and technical solutions. Infrastructure projects like CLARIN actively promote the use of standard formats[1]. However, in practice some standards are available which apparently have high potential but too little actual distribution and application. A reason for this discrepancy between theory and practice can be seen for instance in too much flexibility incorporated in standards, too little knowledge about them in the community, or a limited number of existing applications for data created in adherence with the respective standard.

In the project INEL („Grammatical Descriptions, Corpora and Language Technology for Indigenous Northern Eurasian Languages“)[2] (Arkhipov and Däbritz, 2018) the use of standard formats is especially important because of the long project duration (2016-2033) and the scarcity of the data which covers small, endangered, and under-resourced languages. One of the project's aims is the creation and distribution of language resources, e.g. linguistic corpora. Up until now, corpora and other resources for the languages Dolgan, Kamas, and Selkup have been created and published, and currently work continues on Selkup and Evenki. Several more languages and varieties will be dealt with in the remaining runtime of the project. In addition to the broad coverage and big amount of data, all of the created language corpora include transcriptions of recordings and manuscripts as well as several additional levels of analysis like translations and linguistic annotations, i.e. they come with a rather high degree of complexity.

The following sections focus on the measures taken in the project INEL in order to implement and apply a TEI format that adheres to the ISO standard for the transcription of spoken language. The application of the standard itself is presented and interdisciplinary use cases are discussed, whose presence and demonstration is essential to the beneficial use of the standard, ranging from archiving to search and analysis frameworks.

2 ISO standard for TEI Transcriptions of Spoken Language

With the aim of standard compliance (and integration into the CLARIN-D infrastructure) the project INEL chose a subset of the TEI format as transcription format for the publication and distribution of its

[1] e.g. https://www.clarin.eu/content/standards-and-formats

[2] https://inel.corpora.uni-hamburg.de/

resources. The format follows the P5 guidelines[3] of the Text Encoding Initiative[4] and represents a highly flexible XML format which can be customized on different levels and areas. The high level of flexibility offered by the format (in combination with complex data to be encoded) calls for further customization to allow for specific applications of the created data.

Since the resources in INEL consist of transcriptions of spoken language data, TEI for Speech Representation and the TEI module for spoken language are used. A striking aspect in this context is that the used format adheres to the standard ISO 24624:2016 Language resource management – Transcription of spoken language (ISO/TC 37/SC 4, 2016), which has for instance already been discussed in Fisseni and Schmidt (2019), Schmidt et al. (2017) and Schmidt (2011).

Aside from general benefits that come with standard compliance, the tool-independence of the format allows for the use of various software for the different processing steps of the resource. Another specific benefit lies in the precise modelling enabled by the standard. Challenges of working with the standard format in the project mainly lie in the potential of using the created data in further applications on the one hand and dealing with the creation of the format in absence of a tool native to the standard on the other hand.

2.1 Language Resources in the INEL project

In the INEL project the transcription data is initially created using FLEx[5], which is particularly useful for linguistic glossing. Depending on the data there can be a preliminary step of creating transcripts in ELAN [6] first. The transcriptions are based on audio, video and manuscript data. They are converted into the EXMARaLDA[7] (Schmidt and Wörner, 2014) format for establishing further time alignment, adding annotations and translations as well as performing search, curation and organization of metadata. The need for using multiple tools during the creation of the resource is caused by the different necessary processing steps. As there is no tool that incorporates every functionality needed, the best suited, specialized tool is used for each specific task.

2.2 Applying the ISO standard TEI to the INEL resources

While the EXMARaLDA system provides a solid base for the creation and editing of the resources, its transcription data model is not entirely suitable for the explicit modelling of transcriptions in INEL. Generally, in the EXMARaLDA data model a transcription tier is aligned to the timeline via start and end points. Annotations then are linked to a transcription tier and point to the corresponding start and end points of the transcription elements. The glossing in the INEL data is carried out using annotations based on linguistic segments (tokens). However, segment relations are not a part of the EXMARaLDA transcription data model as it only allows for time-based annotations. Working with the INEL transcriptions in EXMARaLDA is facilitated by matching the time-based start and end points with token boundaries and thus implicitly keeping the segment-relations.

Another phenomenon that cannot be modelled adequately in the EXMARaLDA format is the transcription of words as well as morphemes. In the case of glossing in INEL there are annotations relating to morpheme-based tokens. To deal with this restriction a conversion into the TEI format following the ISO standard for transcription of spoken language is carried out, which makes the implicit token-based annotations explicit and performs a tokenization of the morpheme segments for the glossing. Based on an EXMARaLDA-native export facility an INEL-specific converter was created to cater for the additional processing steps needed for the INEL format/data model. While there already are multiple options and tools available to export ISO/TEI from EXMARaLDA directly, e.g. TEI Drop or the Partitur-Editor export, because of the time-based EXMARaLDA data model, these only export the ISO/TEI format in a purely time-based fashion. The needed export, however, has to precisely represent segment-based INEL transcriptions, so that an adaption and the creation of another converter was unavoidable. With regard to

[3] `https://tei-c.org/guidelines/p5/`
[4] `https://tei-c.org/`
[5] FieldWorks Language Explorer: `https://software.sil.org/fieldworks/`
[6] `https://tla.mpi.nl/tools/tla-tools/elan/`
[7] `https://exmaralda.org/de/`

the existing ISO/TEI-Spoken format for transcription a small extension had to be made in that annotation elements (spans) also refer to other tokenized elements (spans as well) using the ID/IDREF mechanism.[8].

To sum up, the workflow in INEL consists of a pipeline through different stages using several specialized tools ending with the export into the standard ISO/TEI format, which can then be used for further processing, analysis and distribution of the created resources.

3 Use Cases of the resources of the INEL project

The potential use of a standard transcription format touches all areas which merely every transcription format can be concerned with. The difference between standard and non-standard formats mainly lies in the ease of applicability, interoperability, and long-term accessibility. In the following sections multiple application scenarios are presented to demonstrate how the ISO/TEI-Spoken format can be used in different contexts. The state as an ISO standard makes it very suitable for archiving, which is especially crucial for long-term projects that not only aim at preserving data but also continuously and/or periodically work with the data and thus have to ensure a high degree of availability. In addition, it can also very well serve as a base for advanced search and analysis frameworks and further processing.

3.1 Archiving and dissemination

Up until now four linguistic corpora for the languages Dolgan (Däbritz et al., 2019), Kamas (Gusev et al., 2019), and Selkup (Brykina et al., 2018) have been published by the INEL project. It is planned to create corpora for at least seven other languages and varieties until the end of the project in 2033.

The published corpora were ingested into the HZSK Repository[9] which is integrated into the CLARIN infrastructure and uses well-established formats for representing and delivering metadata, e.g. DublinCore, OLAC, and CMDI. The repository mainly contains spoken language corpora similar to the INEL corpora. The main work for their creation is normally done in the EXMARaLDA system, although as mentioned above other transcription tools can also be involved in the creation process. The main transcription format is the EXMARaLDA Basic Transcription (EXB) format, an interoperable XML format[10].

With the ISO standard for the transcription of spoken language available, it was decided to publish ISO/TEI-Spoken[11] alongside the EXMARaLDA transcriptions because the standard is well-documented and tool-independent, and hopefully adds to the accessibility of the data in the long run. The published corpora can also be found in the CLARIN Virtual Language Observatory, e.g. under the keyword "INEL"[12]. A striking challenge, however, persists in the lack of distribution and actual use in the community.

3.2 Analysis and search

To further analyze and search the created resources, the ISO/TEI-Spoken format can be used directly as input or be adapted for use in various existing tools. Depending on the tool the findings or exploration means can be useful for different fields and interdisciplinary research.

For instance, the use of the ISO/TEI-Spoken format in automatic processing with webservices was discussed before by Fisseni and Schmidt (2019) and Schmidt et al. (2017). In Weblicht, the service-oriented architecture for the automatic annotation of linguistic corpora in CLARIN, there is a prototype for a converter[13] from ISO/TEI-Spoken to the Weblicht exchange format TCF and vice versa. Even though this prototype did not reach more than development status, it can in principle be used to process

[8]The refined XML Schema used in the INEL project can be found at `https://gitlab.rrz.uni-hamburg.de/hzsk-open-access/hzsk-corpus-services-release/-/blob/release/src/main/java/de/uni_hamburg/corpora/conversion/resources/xsd/morphemebasedISOTEI.xsd`

[9]`https://corpora.uni-hamburg.de/hzsk/de/repository-search`

[10]Media type of the EXMARaLDA Basic Transcription format:
application/xml;format-variant=exmaralda-exb

[11]Media type of the ISO/TEI-Spoken XML format:
application/tei+xml;format-variant=tei-iso-spoken

[12]`https://vlo.clarin.eu/search?1&q=INEL`

[13]`http://hdl.handle.net/11022/0000-0001-B545-5`

TEI exports of INEL data. However, at the moment firstly the conversion result is quite underspecified, and secondly there is no webservice available that could further process the resulting TCF for the INEL languages (e.g. Dolgan, Kamas, Selkup) or at least unknown languages. In TEILicht[14], several other webservices are already available for directly processing ISO/TEI-Spoken transcripts. In the future, INEL considers contributing webservices which directly process ISO/TEI-Spoken. Several candidates for such services are conceivable, for instance performing different curation tasks like checking and automatically fixing errors in transcriptions or visualizing contents.

Different tools for content search of linguistic resources were evaluated and/or implemented for the INEL resources in the ISO/TEI-Spoken format, e.g. TsaKorpus[15], a linguistic corpus search platform with media support that expects JSON files as input for its search interface. These input files can be created from various base formats. In the beginning of the evaluation the EXB format was used, which was problematic since the imprecision of the segmentation in the files was also present in the search. Now the ISO/TEI-Spoken files are the base format for the search. Further information on how the import is done is available in Arkhangelskiy et al. (2019).[16]

Another search platform is ANNIS[17], a web based search and visualization architecture for complex multi-layer linguistic corpora with diverse types of annotation. It is suitable for the multiple layers of annotations present in the INEL data. After an evaluation using the EXB format as well, it became clear that the ISO/TEI-Spoken format needs to be used for the desired results. While issues of the correct fine-grained mapping of the (multi-speaker and spoken) resources still need to be overcome, an ANNIS instance making the INEL corpora available will be provided in the future.

The MTAS[18] multi-tier annotation search also offers an ISO/TEI parser for the indexing process and was evaluated for the INEL resources. While the evaluation of different queries was successful, the lack of an existing graphical user interface impedes the publication of an MTAS-based search interface. However, there are plans for a future adaptation including a usable interface.

One type of use cases that still needs to be evaluated and explored specifically concerns morpheme based search facilities, and the MTAS search functionality looks promising for a further development of such a tokenized search. Another use case to be explored in the future is the CLARIN Federated Content Search, especially for providing an interdisciplinary or low-threshold access to the resources.

4 Conclusion

The usage of a standard poses specific challenges on the one hand, but also yields evident benefits on the other hand. Several existing and prospective application scenarios were shown in this contribution. It should have become apparent that especially a wide-ranged interdisciplinary further use of the created data encourages and facilitates the use of standards in general, and the discussed ISO/TEI-Spoken format specifically. In the end, the approach to settle for a standard format for dissemination and long-term archiving of the resources and applying the data to as many existing tools as possible definitely proves valuable.

Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

[14] https://clarin.ids-mannheim.de/teilicht/

[15] https://bitbucket.org/tsakorpus/tsakorpus

[16] The indexed INEL resources are freely available for searching at https://www.slm.uni-hamburg.de/inel/portal.html

[17] https://corpus-tools.org/annis/

[18] https://meertensinstituut.github.io/mtas/

References

Timofey Arkhangelskiy, Anne Ferger, and Hanna Hedeland. 2019. Uralic multimedia corpora: Iso/tei corpus data in the project inel. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 115–124.

Alexandre Arkhipov and Chris Lasse Däbritz. 2018. Hamburg corpora for indigenous northern eurasian languages. *Tomsk Journal of Linguistics and Anthropology*, 21(3):9–18.

Maria Brykina, Svetlana Orlova, and Beáta Wagner-Nagy. 2018. INEL Selkup Corpus. Version 0.1. Publication date 2018-12-31. http://hdl.handle.net/11022/0000-0007-CAE5-3. Archived in Hamburger Zentrum für Sprachkorpora. In Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger, Daniel Jettka, and Timm Lehmberg, editors, *The INEL corpora of indigenous Northern Eurasian languages*.

Chris Lasse Däbritz, Nina Kudryakova, and Eugénie Stapert. 2019. INEL Dolgan Corpus. Version 1.0. Publication date 2019-08-31. http://hdl.handle.net/11022/0000-0007-CAE7-1. Archived in Hamburger Zentrum für Sprachkorpora. In Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger, Daniel Jettka, and Timm Lehmberg, editors, *The INEL corpora of indigenous Northern Eurasian languages*.

Bernhard Fisseni and Thomas Schmidt. 2019. CLARIN Web Services for TEI-annotated Transcripts of Spoken Language. *Proceedings of the CLARIN Annual Conference 2019, CLARIN ERIC: Leipzig*, page 36–39.

Valentin Gusev, Tiina Klooster, and Beáta Wagner-Nagy. 2019. INEL Kamas Corpus. Version 1.0. Publication date 2019-12-15. http://hdl.handle.net/11022/0000-0007-DA6E-9. Archived in Hamburger Zentrum für Sprachkorpora. In Beáta Wagner-Nagy, Alexandre Arkhipov, Anne Ferger, Daniel Jettka, and Timm Lehmberg, editors, *The INEL corpora of indigenous Northern Eurasian languages*.

ISO/TC 37/SC 4. 2016. Language resource management – Transcription of spoken language. Standard ISO 24624:2016, International Organization for Standardization, Geneva, CH.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Hanna Hedeland, and Daniel Jettka. 2017. Conversion and annotation web services for spoken language data in CLARIN. *Selected papers from the CLARIN annual conf. 2016*, page 113–130.

Thomas Schmidt. 2011. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*.

Evaluating and Assuring Research Data Quality for Audiovisual Annotated Language Data

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

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

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

Abstract

This paper presents the QUEST project and describes concepts and tools that are being developed within its framework. The goal of the project is to establish quality criteria and curation criteria for annotated audiovisual language data. Building on existing resources developed by the participating institutions earlier, QUEST develops tools that could be used to facilitate and verify adherence to these criteria. An important focus of the project is making these tools accessible for researchers without substantial technical background and helping them produce high-quality data. The main tools we intend to provide are the depositors' questionnaire and automatic quality assurance, both developed as web applications. They are accompanied by a Knowledge base, which will contain recommendations and descriptions of best practices established in the course of the project. Conceptually, we split linguistic data into three resource classes (data deposits, collections and corpora). The class of a resource defines the strictness of the quality assurance it should undergo. This division is introduced so that too strict quality criteria do not prevent researchers from depositing their data.

1 Introduction

The QUEST[1] project is one of twelve projects recently funded by the German Federal Ministry of Education and Research across all disciplines with the aim of enhancing research data quality and re-use. As the full title, "Quest: Quality - Established: Testing and application of curation criteria and quality standards for audiovisual, annotated language data", suggests, the focus is on one particular resource type, for which reliable quality standards and curation criteria will be developed. The project, which runs from 2019 to 2022, is based on the existing cooperation within the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[2] (Hedeland et al., 2018), comprising the Data Center for the Humanities (DCH)[3] and the Department of Linguistics (IfL)[4] (both Cologne), the Endangered Language Archive (ELAR)[5] and the SOAS World Languages Institute (SWLI)[6] (both London), the Hamburg Centre for Language Corpora (HZSK)[7] and the long-term project INEL[8] (both Hamburg)

[1] https://www.slm.uni-hamburg.de/en/ifuu/forschung/forschungsprojekte/quest.html
[2] http://ckld.uni-koeln.de/
[3] https://dch.phil-fak.uni-koeln.de/
[4] https://ifl.phil-fak.uni-koeln.de/en/
[5] https://www.soas.ac.uk/elar/
[6] https://www.soas.ac.uk/world-languages-institute/
[7] https://corpora.uni-hamburg.de/hzsk/en
[8] https://www.slm.uni-hamburg.de/inel/

and the Leibniz Centre General Linguistics[9] (ZAS, Berlin). For the QUEST project, the CKLD members have been joined by the German Sign Language Corpus project (DGS-Korpus)[10] in Hamburg and the Archive for Spoken German (AGD)[11] at the Institute for German Language (IDS) in Mannheim, who bring in their respective expertise. With the focus on annotated audiovisual language data, the aim of the project is twofold. On the one hand, it is to develop generic quality criteria valid regardless of intended usage scenarios. On the other hand, it aims to establish specific curation criteria tailored to certain re-use scenarios related to individual disciplines and/or research methods. To enable researchers to adhere to such criteria, these must be both adequate and not conflicting with research. Additionally, there must be comprehensive support for researchers with little technical background in applying them to their data, which is another important part of the project's goals.

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

2 Background

The conceptual parts of QUEST regarding the definition of criteria draw on the expertise gathered within all project members' institutions and other relevant organisations. For the implementation of the quality assurance system, previous efforts by the data centres AGD (the Archive for Spoken German) and the HZSK (the Hamburg Centre for Language Corpora), which are both CLARIN B Centres, play a major role. One such existing resource we build upon is the assessment guidelines for legacy data (Schmidt et al., 2013), which were developed to set minimal standards for data deposits and make decisions regarding data curation transparent. The need to handle the increasing amount of incoming resources with more efficiency and transparency at the Hamburg Centre for Language Corpora led to the development of another resource, the HZSK Corpus Services (Hedeland and Ferger, 2020). The HZSK Corpus Services are a complex framework for data curation and quality control, which are based on the EXMARaLDA system (Schmidt and Wörner, 2014) and are currently enabling efficient collaborative resource curation at the HZSK and within the INEL long-term project, which is based on the HZSK infrastructure. Other relevant approaches not part of the QUEST project include the "Open Source analogy for research data curation"[12] applied in the collaborative workflows of the Cross-Linguistic Linked Data (CLLD) project (Forkel, 2015), the work on continuous quality control and reproducibility for other resource types within the CONQUAIRE (Continuous quality control for research data to ensure reproducibility) project (Cimiano et al., 2015) and, to some extent, the DoorKeeper functionality of the FLAT repository at the Max Plank Institute for Psycholinguistics (MPI) in Nijmegen (Trilsbeek and Windhouwer, 2016), which is focused on archivability rather than content-related resource quality or reproducibility.

3 Resource Types, Data Formats and Curation Levels

The aim of the QUEST project is not to standardize the creation of audiovisual language resources but rather to take stock of the existing heterogeneity and promote such standards and formats in use that lend themselves to quality control. Another important aim is to find means to implement such functionality. A first step is to review and describe variation in existing resources both on the macro-level, i.e. resource structure and the involved data types, and on the micro-level, due to various tier structures, annotation schemes and transcription conventions. Following an inventory of QUEST associated and other relevant (CLARIN) data centres, an initial set of linguistically relevant data types based on their role within a resource was defined as the basis for meaningful recommendations on file formats. This set includes audio and video recordings, transcription/annotation data, lexical databases, additional relevant written or image material, contextual (meta)data on sessions and participants, documentation, catalogue and detailed metadata, and settings files. For generic data types such as audio, video, image and unstructured text files used for documentation there is little controversy regarding good practices for archival formats.

[9] https://www.leibniz-zas.de/en/
[10] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html
[11] http://agd.ids-mannheim.de/index_en.shtml
[12] https://clld.org/2015/02/03/open-source-research-data.html

However, for the file formats used for transcription/annotation data and contextual data, the situation is far more complex.

While a few widely used and interoperable formats (such as ELAN (Sloetjes, 2014) or EXMARaLDA) are accepted across all centres, the level of structuredness and machine readability and comprehensibility of resources created with these formats differs widely depending on the research methods employed, especially as qualitative approaches do not rely on machine-readable data. While the original research might not profit from structured and machine-understandable data, discoverability and future re-use scenarios depend on these aspects. Since data curation is however a very costly endeavour and only partly possible for orphaned resources, a division in three resources classes was developed to avoid strict quality criteria becoming obstacles for depositing valuable data:

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

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

4 Quality Control for QUEST Data Centres and Users

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

4.1 A Planning and Evaluation Tool for Depositors

The content of the depositors' questionnaire (Schmidt et al., 2013) was migrated to a new technical solution and extended according to the QUEST context. The questionnaire is now implemented as a web application and serves as the initial step of the quality control pipeline. Unlike the original questionnaire, the updated one can be used in two scenarios, which contain different sets of questions. In the first scenario, the user is planning a project and does not have the actual data at hand. In this case, they answer questions regarding their prospective data, e.g. whether they are going to have morphological annotation. At the end, the questionnaire generates templates tailored to the user's needs that can be used throughout the project. At the moment, supported formats are ELAN template files and EXMARaLDA stylesheets. Both can be used for creating new empty annotations in the respective software. This ensures that the data will have consistent annotation, thus reducing curation workload after the project is complete. In the second scenario, it is assumed the data has already been prepared, and the user would like either to deposit it to a QUEST center, or just to make sure it conforms to the basic quality requirements. In this scenario, the distinction between the three resource classes described in 3 is made. Depending on the resource class selected by the depositor at the beginning of the questionnaire, some of the questions may be skipped. If the user's responses indicate problems that prevent their data from undergoing further quality control, such as lack of informed consent, the questionnaire app lists them together with tips that could help resolve them. If no such problem is found, the user receives a settings file with the summary of their responses, which can later be submitted to the second stage of quality control. These settings turn certain checks on or off, as well as provide parameter values (such as transcription tier name) to some checks.

4.2 A Flexible Quality Control Framework

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

First direction is the usability. Corpus Services are a Java application that can only be run from the terminal; additionally, the user must pass dozens of arguments to switch particular tests on or off. In projects working with the software, this is done by using batch scripts customized for individual resources. Since this is beyond limits to most ordinary linguists, a web application was developed to make the testing process accessible to a wider audience. The front end is a web page that allows the user to upload an archive with the corpus to be tested, along with a settings file generated by the questionnaire (section 4.1). The back end unpacks the archive in a temporary folder on the server and runs Corpus Services with arguments defined in the settings file. After the test is complete (which may take minutes or even hours), the corpus files are removed from the server. The HTML report generated by Corpus Services is then sent to the user via email. It can also be accessed afterwards on the server through a unique URL generated at upload time and shown to the user. Although this solution cannot be applied to corpora that are too large to be uploaded, we believe it will still cover the majority of cases.

Second, the contents of the framework is extended according to the QUEST context, since currently, only EXMARaLDA data can be validated. First and foremost, this means adding the ability to process the EAF format of the ELAN software used by the centres in London, Cologne and Berlin, and preferably also the FOLKER format used at the Archive for Spoken German and, possibly, other formats. Also, many more checks/services should be added for generic and specific criteria developed within the QUEST project. This part of the extension is in its initial stage now.

4.3 A Common Knowledge Base

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

5 Outlook

Since common widely accepted recommendations and support in adhering to them are still lacking for researchers working with audiovisual language data, the work within the QUEST project can hopefully gain impact and applicability beyond original QUEST centres through the CLARIN Knowledge Sharing Infrastructure connection. It might also provide valuable input for the creation of Domain Data Protocols for audiovisual annotated language resources as suggested by Science Europe (Science Europe, 2018), which might be a way of providing quality criteria to users in a transparent and applicable manner. Providing various diagnostic tests for audiovisual resources that can be used at deposit but also during resource creation to external projects will allow these to prepare for data deposit and make this process more transparent, resulting in more high quality resources becoming available for interdisciplinary re-use within existing and emerging digital research infrastructures for the humanities and social sciences.

References

Philipp Cimiano, John McCrae, Najko Jahn, Christian Pietsch, Jochen Schirrwagen, Johanna Vompras, and Cord Wiljes. 2015. CONQUAIRE: Continuous quality control for research data to ensure reproducibility: an institutional approach, September.

Robert Forkel. 2015. Cross-Linguistic Linked Data: Dateninfrastruktur für Diversity Linguistics. In *Forschungsdaten in den Geisteswissenschaften (FORGE) 2015, (Hamburg, 5-18 September, 2015)*, pages 10–12, Hamburg.

Hanna Hedeland and Anne Ferger. 2020. Towards continuous quality control for spoken language corpora. *International Journal for Digital Curation*, 15(1).

Hanna Hedeland, Timm Lehmberg, Felix Rau, Sophie Salffner, Mandana Seyfeddinipur, and Andreas Witt. 2018. Introducing the clarin knowledge centre for linguistic diversity and language documentation. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018), 7-12 May 2018, Miyazaki, Japan*, pages 2340 – 2343, Paris, France. European language resources association (ELRA).

Thomas Schmidt and Kai Wörner. 2014. Exmaralda. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2013. Leitfaden zur beurteilung von aufbereitungsaufwand und nachnutzbarkeit von korpora gesprochener sprache.

Science Europe. 2018. Science Europe Guidance Document Presenting a Framework for Discipline-specific Research Data Management, January.

Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.

Paul Trilsbeek and Menzo Windhouwer. 2016. FLAT: A CLARIN-compatible repository solution based on Fedora Commons. In *Proceedings of the CLARIN Annual Conference 2016*. CLARIN ERIC.

Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources

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

Abstract

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

1 Introduction

The successful development of large digital research infrastructures such as CLARIN has enabled the sharing and re-use of language resources across geographic and, partly, disciplinary boundaries. This has led to a shift in focus from the technical means of data sharing towards the data itself and in particular its quality and fitness for re-use. However, while e.g the German Council for Scientific Information Infrastructures (RfII) states in the latest of their recommendations that "securing and improving data quality is a fundamental value of good scientific practice" (RfII, 2020), widely acknowledged and adequate definitions of data quality for the various types of language resources provided through digital infrastructures are still lacking. Generic approaches such as the FAIR Principles (Wilkinson and others, 2016) or even the FAIR Metrics (Wilkinson et al., 2018) do not provide detailed guidance for research data management for specific resource types or research methods related to specific disciplines. The metrics only refer to data formats "recommended by the target research community" and since the metrics are not resource type or discipline specific, it is not possible to formulate more specific criteria for the data within these generic metrics.

Research data quality calls for adequate and comprehensive definitions, but this raises several – often overlooked – fundamental questions. Suitable quality criteria need to be transparent and operationalized, but also reflect the complexity of the subject matter, audiovisual annotated language data. A first step is therefore a review of this resource type, before various approaches to defining data quality criteria can in turn be evaluated in terms of their applicability.

2 Taking Stock of Audiovisual Annotated Language Data Resources

The various resource types subsumed under "audiovisual annotated language resources" are highly heterogeneous but have in common that they comprise several data types and display a complex structure of abstract entities and data objects with different types of relations. A comprehensive description of these resources and the variation within the group is therefore an important first step. This is one goal of the German QUEST project, based on the the existing cooperation of the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD)[1] – including DCH/IfL (Cologne),

[1] http://ckld.uni-koeln.de/

ELAR/SWLI (London), HZSK/INEL (Hamburg) and ZAS (Berlin) – and extended by the German Sign Language Corpus project (Hamburg) and the Archive for Spoken German at IDS (Mannheim), adding their complementary expertise on German data and sign language data, respectively.

The participating data centres partly allow self-deposit of resources with basic requirements on file formats and metadata – ELAR and the Cologne Language Archive (LAC) –, and partly curate resources to comply with corpus data models and achieve data consistency – the AGD and the HZSK. Since the data deposited with the AGD and the HZSK is often from projects working with qualitative methods only, all requirements are not really relevant to the original research, which is obviously reflected in the data to a varying extent. The resources in all four centres differ along several dimensions, which can be described as structural, methodological and content-based heterogeneity.

2.1 Structural Heterogeneity

Abstract data models for language resources such as EXMARaLDA (Schmidt and Wörner, 2014) or the DGD data model(Schmidt et al., 2013a) provide explicit structural requirements on not only the macro-level of abstract entities and data files, but also regarding the micro-level, with consistency in tier structure and content and regarding the identity of speakers. Even without an explicit data model, the resource structure is also defined by contextual data, including structurally relevant entities such as recording sessions, and by metadata on included files and their relations, e.g. IMDI[2] or CMDI[3] metadata. However, not all resources exploit such models or schemes, but simply amount to a set of audio and video recordings and individual transcripts, in some cases with no explicit information on the internal structure available and only minimal metadata. Though resources typically contain the same type of abstract and data objects, the corresponding file formats vary. In particular, they are either unstructured, e.g. as transcription data provided in PDF or plain text format, or structured, e.g. as XML transcription/annotation tool formats.

2.2 Methodological Heterogeneity

Differences on the micro-level are strongly dependent on the research methods employed, especially whether qualitative or quantitative/automatic analysis have been used. Annotation thus range from tags from a controlled scheme added in a systematic and comprehensive way to interpretative free text comments added to relevant parts only. Since transcription conventions capture and stress certain aspects of language, they also differ with respect to units such as utterances or intonation phrases and the amount of linguistic information integrated into the basic transcription. Furthermore, not all transcription and annotation schemes in use lend themselves to automatic syntax checking.

2.3 Content-related Heterogeneity

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

3 Approaches to Data Quality and Possible Applicability for Language Resources

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

3.1 Generic Approaches to Data Quality

Generic approaches do not restrict the types of data they are applicable too and thus recommendations remain general and abstract. (Wang and Strong, 1996) distinguish fundamental dimensions: intrinsic,

[2] https://archive.mpi.nl/forums/t/imdi-metadata-information/2639/2

[3] https://www.clarin.eu/cmdi

contextual, representational and accessibility data quality, pertaining to the data itself, a particular usage context, and the systems providing data, respectively. This distinction between inherent and system-dependent data quality is also reflected in ISO/IEC 25012 - The Data Quality Model[4]. The W3C provide relevant input in their Best Practices for Data on the Web[5], both regarding the recommendations and the system used to disseminate them. However, these generic approaches do not provide directly applicable resource specific recommendations.

3.2 Approaches to Research Data Quality

Today, for research data to be FAIR is the main requirement. Even though the FAIR metrics aim to operationalize the well-known principles, they also only refer to community-specific standards. The FAIRification process (Jacobsen et al., 2020) also requires resource type specific requirements and workflows, but is a starting point to redefine data curation processes in line with FAIR concepts.

3.3 Resource Type Specific Approaches to Data Quality

Within CLARIN, there is work in progress to collect recommendations from all CLARIN B centres on standards and formats accepted for deposit[6]. Apart from the participants of the QUEST project, some centres providing detailed recommendations for audiovisual data are e.g. The Language Archive at the MPI in Nijmegen[7] and the Bavarian Archive for Speech Signals[8]. Furthermore, the German funder DFG has published recommendations for technical standards[9] collected through discussions within the relevant research communities. And still highly relevant after almost twenty years, (Bird and Simons, 2003) have described several aspects relevant for the long-time preservation and re-use of language documentation data. These are valuable resources for definitions of data quality.

An important aspect which is beyond the scope of technical recommendations on standards and formats, but at the same time must be considered at all times, is the quality of research data as an artefact of research, which can only be as good as the research (and vice versa).

4 Step One: Defining Classes of Audiovisual Resources

Considering the heterogeneity, it would be inappropriate to measure quality without regarding the conscious choices and trade-offs made by researchers leading to the encountered differences. The AGD and the HZSK have defined guidelines for deciding whether to perform data curation (Schmidt et al., 2013b) and while curation of deposited data increases the re-use potential, with increasing deposit numbers, the task becomes impossible. By allowing controlled variation, re-users know what to expect from resources data and more adequate goals for evaluation and curation can be defined. While the focus here is on audiovisual data, the following categories could also be applied to written language resources.

4.1 Deposits

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

4.2 Collections

On the next level, Collections comply with additional requirements on the macro-level, including the completeness and consistency of metadata and relations between all resource parts. The completeness of metadata only pertain to standardized cataloguing metadata describing the linguistic resource and its

[4] https://www.iso.org/standard/35736.html
[5] https://www.w3.org/TR/dwbp/
[6] cf. https://www.clarin.eu/content/standards-and-formats
[7] https://archive.mpi.nl/tla/accepted-file-formats
[8] https://www.phonetik.uni-muenchen.de/Bas/BasInfoStandardsTemplateseng.html
[9] https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf

provenance to make the data comprehensible, since contextual data, e.g. information on participants, can not be standardized without interfering with research design. The language data of Collections is provided in various unstructured text formats suitable for human manual analysis.

4.3 Corpora

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

5 Step Two: Data Curation as FAIRfication

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

The ISO standard for Transcription of Spoken Language[10] provides more semantic information on units and information types as part of the underlying data model than most widely used formats for transcription/annotation data, which do not define the notation of e.g. participants' contributions, noise or pauses. As the standard was developed with this idea in mind, conversion would be one step towards semantic interoperability, though still there is no designated method to include machine-readable references for tiers or individual annotations in this TEI-based format. Additional conventions would allow for a proper definition of the semantics of individual data sets and increase the options for re-use, especially within NLP contexts.

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

While the aspects of FAIRification (from structuredness to semantic enrichment and linking) are generic, data quality is to a great extent a question of the data being fit for particular purposes or usage scenarios – and not all usage scenarios improve by using more structured data.

Since it is not feasible for research projects creating language resources to consider all possible re-use scenarios, explicit and formalized definitions of re-use scenarios would allow projects to comply with specific re-use scenarios. Re-users would also be able to recognize whether the data is suitable for their purpose, which is often difficult to tell today, especially in the case of interdisciplinary re-use, e.g. between linguistics and education sciences, partly also due to the use of different terminology.

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

[10] `https://www.iso.org/standard/37338.html`

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

7 Outlook

Though seemingly trivial, fundamental questions regarding the structure and content of annotated audiovisual language resources created as research data within various disciplines have yet to be thoroughly discussed and answered. The characteristics of such resources need to be systematically described in order to define suitable criteria for data quality. Providing generic quality criteria applicable to resources with various levels of curation and structuredness is one aim of the QUEST project, another is to provide additional criteria for formalized re-use scenarios. To allow data creators to comply with these criteria, the project will also provide software solutions to evaluate various types of resources accordingly and preferably continuously during data creation. In combination, the definitions and evaluation mechanisms developed within the QUEST project will hopefully make data depositing and re-use more transparent and fruitful within and across disciplines.

References

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582, September.

Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, and Mark Thompson. 2020. A generic workflow for the data FAIRification process. *Data Intelligence*, 2:56–65.

RfII. 2020. The Data Quality Challenge. Recommendations for Sustainable Research in the Digital Turn.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In Ulrike Gut Jacques Durand and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Thomas Schmidt, Sylvia Dickgießer, and Joachim Gasch. 2013a. Die datenbank für gesprochenes deutsch - DGD2.

Thomas Schmidt, Kai Wörner, Hanna Hedeland, and Timm Lehmberg. 2013b. Leitfaden zur beurteilung von aufbereitungsaufwand und nachnutzbarkeit von korpora gesprochener sprache.

Han Sloetjes. 2014. ELAN: Multimedia annotation application. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *Handbook on Corpus Phonology*, pages 305–320. Oxford University Press.

Kilu von Prince and Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France, May. European Language Resources Association.

Richard Y. Wang and Diane M. Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33.

Mark D. Wilkinson et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018–, March.

Mark Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5:180118, 06.

Towards an Interdisciplinary Annotation Framework: Combining NLP and Expertise in Humanities

Laska Laskova
AIaLT
IICT-BAS, Bulgaria

Petya Osenova
AIaLT
IICT-BAS, Bulgaria

Kiril Simov
AIaLT
IICT-BAS, Bulgaria

{laska|petya|kivs}@bultreebank.org

Abstract

The paper describes the initial steps in creating an annotation framework that would incorporate the knowledge coming from Bulgarian corpora, lexicons, linguistic analyzers with the expert knowledge coming from specialists in History, Diachronic Linguistics, Iconography. The proposed framework relies on the INCEpTION system, CIDOC CRM ontology and FrameNet. Here the following steps are described: workflow, guideline principles and challenges. The domain focus is History. The ultimate goal is to provide enough manual and verified expert data for constructing a Bulgaria-centered knowledge graph in addition to the information coming from Wikipedia, Wikidata, Geonames. The annotations will be used also for training automatic semantic processors and linkers.

1 Introduction

Within CLaDA-BG, our goal is to construct a Bulgaria-centric Knowledge Graph (BGKG) which represents facts related to Bulgarian people, locations, events, and organizations. We follow the approach of Fokkens et al. (2017) in applying NLP techniques to domain specific texts in order to extract knowledge for BGKG. There are many initiatives that use NLP techniques for automatic mapping of biographies to ontologies and Linked Open Data (LOD). In addition to Fokkens et al. (2017) see for example Hyvönen et al. (2014) and Tuominen et al. (2018) for mapping with ontologies, among others. Our aim is to arrange the knowledge in BGKG with respect to the events in which different objects (people, locations, organizations, artefacts, etc.) participate, similarly to Rospocher et al. (2016). In the actual construction of BGKG, we envisage to reuse all the available knowledge in the existing knowledge graphs like Bulgarian version of DBPedia[1], Wikidata[2], Bulgarian Wikipedia, etc. in order to cover a wide range of knowledge. But our main source of information are expert scientific publications and primary scientific data like archive collections, historical publications, etc. Compared to the above-mentioned initiatives for languages like Dutch or Finnish (not to mention English), there are insufficient language resources and language technologies for the automatic construction of a Bulgaria-centric knowledge graph. For this reason, we decided to adapt the available language resources and technology by adding focused semantic annotation of domain specific documents.

In this work, we report on the design and implementation of the annotation schema as tuned to our goals. Our motivation is to select the right level of granularity which would allow an efficient annotation process with minimal level of ambiguities for annotators while keeping the precision of the annotation as high as possible and the loss of useful information as less as possible. Thus, we started with an event-based ontology, namely CIDOC-CRM.[3] This ontology was designed to represent data from domains that we would like to cover within BGKG, namely Social Sciences and History where biographies play a central role. In addition, we use FrameNet as a provider of relevant situation descriptions with their corresponding participants. Last, but not least, Named Entities (NEs) annotation level is provided. Note

[1] http://bg.dbpedia.org/

[2] https://www.wikidata.org/wiki/Wikidata:Mai_Page

[3] http://www.cidoc-crm.org/

that our NE typology is wider and covers more named elements. In addition to *Person*, *Location*, and *Organization*, it also includes *Sum*, *Time*, *Etnonym*, etc. The ultimate goal is to ensure linking of Named Entities, events and participants in domain specific texts to concepts in the ontology, WordNet, and Wikipedia.

2 Annotation Tool and Workflow

We needed a web-based tool that is suitable for both annotation (with curation) and linking (to ontologies), and also easy enough for domain experts to work with. Several systems were considered, such as BRAT[4], SALT[5], and WebAnno[6]. Despite the fact that all the mentioned systems are appropriate, finally the system INCEpTION[7] was selected - Klie et al. (2018). The reason is as follows: it was developed on the basis of WebAnno, and our team had previous working experience with WebAnno for a coreference annotation task. INCEpTION provides a friendly visualization and a possibility for level-based annotation with pre-defined information.

Our strategy is as follows. First, several meetings were organized within the NLP group for initial acquaintance with the tool and training – how to upload a document, how to download it, how to tune the settings, how to annotate, how to facilitate annotation when there is repeated information, etc. Guidelines were developed gradually starting from basic phases and moving to more complex ones. After that, training started with our colleagues from two institutes of Bulgarian Academy of Sciences (BAS), specialists in Balkan history and investigators of life and works of Cyril and Methodius. First, annotation guidelines that reflect the specifics of biographical texts were created. They were applied to documents related to Bulgarian donors for education from 1919 to 1945. Later on, annotation started on domain specific texts in order to re-design the initial guidelines with respect to the needs of the partners. Questions-and-Answers sessions were regularly organized in which the annotators shared their uncertainties and difficulties, and received feedback about the curated files. The annotated and linked documents will be made available through the CLaDA-BG repository which is a LINDAT-based one[8]. More details on the guidelines evolution and challenges are given in Sections 3 and 4 respectively.

3 Annotation Guidelines

The first version of the guidelines was tailored for the annotation of 37 documents (276 KB in total) on donor acts in Bulgaria (1919-1945)[9], structured in the following manner: a short biographical sketch, description of the donations, and, in some cases, a copy of a donation document, a letter or an excerpt from a will. The initial annotation task group included eight members, four of them experienced annotators, i.e. people who have a lot of previous experience as annotators in various projects.

First, we focused on semantic typing with only one annotation level specified, that of Named Entities. We introduced 11 types: EVT (event), JUR (juridical), LOC (location), LOC-GPE (geo-political location), ORG (organization), PER (person), PER-GPE (group/person, part of GPE), PRO (product), REF (reference), SUM (sum of money), TIME (dates, years). While the last two are not rigid designators in the sense of Kripke (1971), they are often considered a part of the standard Named-Entity Recognition task because of their "practical importance" (Nouvel et al., 2016). Then, a second, semantic role level, based on the several pivotal types of life events, was added. It includes schemes for 9 states-of-affairs: *Relatives, Birth, Death, Living, Moving, Education, Work, Relation, Charity*. These schemes are based on the corresponding FrameNet[10] entries and each of them is associated with a list of typical elements or participants (roles). For example, whenever the annotator tags a word or a phrase that evokes *Birth* scheme, s/he is offered a preset number of roles: parent (person), born (person), place, time. The set

[4] http://brat.nlplab.org/index.html
[5] https://www.saltsoftware.com/
[6] https://webanno.github.io/webanno/
[7] https://inception-project.github.io/
[8] https://lindat.mff.cuni.cz/en/about-lindat-clarin
[9] National project DFNI-K02/12 (12.12.2014) Culture of giving in the sphere of education: social, institutional and personality dimensions.
[10] https://framenet.icsi.berkeley.edu/fndrupal/

of participants was decided based on their salience for the enrichment of the BGKG. Participants that are peripheral or non-core from a linguistic perspective, such as time and place, are in fact crucial for the complete description of any event. On the other hand, the text does not always provide information for all the core elements. For example, when the annotator chooses the *Education* tag, s/he is offered the following list of participants: learner (person), teacher (person), institution (organization), speciality (object), level (quality), place, time (or more specifically, start, end, duration). Figure 1 shows an annotated snippet[11], where the verb завършва, zavarshva (graduates) is tagged as *Education*. All elements of the *Education* scheme mentioned in the text are connected to завършва with a labeled relation: learner (Ганчо Ценов, Gancho Tsenov), speciality (история, history), end (1894 година, year 1894), and organization (Софийския университет, Sofia University). Some elements, for example the name of the teacher or the year of enrollment, are not indicated in the document, others like place or education level are inferrable.

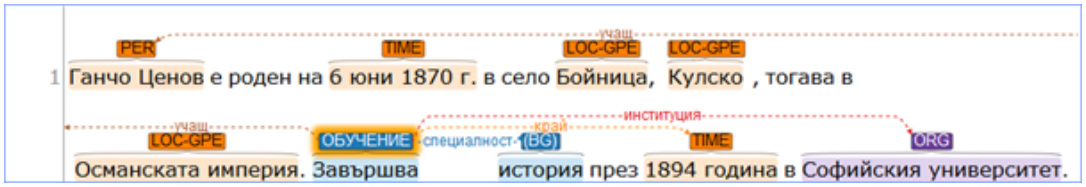


Figure 1: Annotation example. The word Завършва that invokes the scheme *Education* (ОБУЧЕНИЕ in light blue), is connected with labeled dotted arrows to the linguistic expressions that denote four of the elements typical for this situation. To the left: учащ (learner) that goes to PER. To the right: специалност (specialty) which in this case is history; край (end) which denotes the TIME of completion; институция (organization) which goes to ORG.

While *Birth* and *Education* schemes are typical for most biographies, *Charity* is a situation specific to the documents in the corpus. *Charity* has the following list of elements: donor (person), recipient (person or organization), mediator (person or organization), time, validity (how much time the recipient benefits from the charity act), manner (secretly, under specific condition), goal (the intended situation resulting from the act of donation).

We measured the inter-annotator agreement with Cohen's kappa at all active levels of annotation. Thus, at the NE annotation level, the highest score is 1 with the involvement of 2 experienced annotators, and the lowest score is 0,87 with the involvement of 1 experienced and 1 non-experienced annotator; for the event annotation, the highest score is again 1 with the involvement of 2 experienced annotators, and the lowest score is 0,91 with the involvement of 2 non-experienced annotators; as for the annotation of participants (roles), the highest score is 1 regardless of the annotators' experience, and the lowest score is 0,87 with the involvement of 1 experienced and 1 non-experienced annotator. Note that after the annotation, a super-annotator checked the results providing the final curation. At this stage it seems that the event annotation is slightly clearer and easier to handle compared to the annotation of NEs and event participants. The fact that there is no apparent correlation between the annotators' experience and the inter-annotator agreement as far as the annotation of roles is concerned, might be connected to the limited number of tags that the annotator is offered by the system once the event type is set.

An important feature of the annotation strategy was adopted after the completion of the initial training. It is the following: the annotator had the freedom to add a new type of NE or a scheme tag, if deemed necessary. In this way the detection of specific terminology was addressed. The experts were given texts from their domain of expertise. From this moment on, the guidelines became domain specific. Current domain texts in history cover various research topics summarized under the title *Thessaloniki and the Bulgarians: History, Memory, Present* and the annotation process is still ongoing.

[11]The translation of the snippet is as follows: Gancho Tsenov was born on June 6, 1870 in Boynitsa village, Kula region, which was at that time within the borders of the Ottoman empire. In 1894, he completed his studies in History at Sofia University.

The tags at this point are matched to various CIDOC-CRM concepts, such as: E39 Actor with subclasses E21 *Person* and E74 *Group* (with a subclass of E40 *Legal Body*); E49 *Time Appellation* (with a subclass E50 *Date*); E52 *Time-Span*; E61 *Time Primitive*; E95 *Spacetime Primitive*; E53 *Place*; E97 *Monetary Amount*; E5 *Event* with subclasses: E63 *Beginning of Existence* (with E65 *Creation*, E66 *Formation*, E67 *Birth*, E81 *Transformation*) and E64 *End of Existence* (with E6 *Destruction*, E68 *Dissolution*, E69 *Death*, E81 *Transformation*).

FrameNet provides more specific relations (properties) compared to the ones offered in the ontology. For example, for E67 *Birth* all the properties are used: P96 *by mother* (gave birth): E21 *Person*; P97 *from father* (was father for): E21 *Person*; P98 *brought into life* (was born): E21 *Person*. In contrast, the scheme *Charity* is based on the frame *Giving* as described in FrameNet[12], with core participants *Donor*, *Theme* and *Recipient*; after that it was mapped to E7 *Activity* from CIDOC-CRM.

Combining the mappings to an ontology and the FrameNet ensures both - reasonable granularity and proper inheritance - in presenting the domain specific events and their participants.

4 Challenges

The challenges we face are mainly in three directions: *system-based*, *representational*, and *content-wise*. The *system-based* ones refer to the constraints of the tool and the depth of our knowledge about it. For example, at the moment there is no direct way of connecting text chunks with ontology concepts. The *representational* problems stem from the fact that for many language structures there are different annotation approaches. Coordinated structures with elliptical material (e.g. “South and North Korea”; “24, 25 and 26 Vasil Levski Str.”, etc.), to mention one example, can be represented by copying the missing parts or by dividing the annotation into several parts, tagged appropriately. Another point of discussion is the criteria by which the length of a NE or a scheme participant should be solved. This issue includes: nested NEs; which elements to be part of a tag; the position of one-level element to another level element, etc.

The *content-wise* challenges refer to: which relations between names and concepts should be made explicit, and which can be derived automatically (e.g. in the phrase “aunt Annie”, should we put the relation *is-a* between the concept “aunt” and the name “Annie” or annotate the two words together and analyse the relation on a later stage?); how many relations should be added, i.e. what degree of density is needed for manual work and what can be performed automatically; using a stacking or an indexing technique for the same situation with different numbers of varying participants; with what types of preprocessing to further facilitate the work of the domain experts, etc.

5 Conclusion

The semantic annotation of texts specific to History and other Humanities-related domains, proved to be a very useful intersection point between people working in Language Technologies and experts in important branches of Social Sciences and Arts. On the one hand, NLP-ers get feedback on coverage and provenance of computationally developed lexicons and tools. On the other hand, experts have the chance to deliver the pieces of knowledge they worked on, in a structured and contextualized way. Thus, the balance has to be found between explicit and implicit information, specialist and non-specialist understanding of various events, important and side facts and opinions.

Our intention is to continue this endeavor in the following directions: developing various domain specific guidelines that reflect the respective domain knowledge; increasing the automatic components as preprocessing (NEs detection and NEs linking) and postprocessing, integrating CIDOC-CRM and LOD (Wikidata, etc.) ontology through direct mappings. During the next phase of annotation to each of the annotations appropriate identifiers will be added. These identifiers will be part of the BGKG representing instances or classes and properties in the related ontologies.

The benefit from the manual annotation of the domain documents will be manifold:

[12] `https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Giving`

1. *annotation guidelines for experts*: we consider annotation, curation and discussion an essential part of guidelines creation and revision;

2. *semantic analyzers* and *knowledge extractors*: it provides the basis for the creation of annotation pipelines for knowledge extraction from historical documents;

3. *methodological insights*: the developed a methodology of combining language and semantic resources and technologies that can be reused in new domains of interest.

We believe that these lessons learnt and to be learnt in our ongoing work are valuable contributions to both CLaDA-BG and CLARIN infrastructures.

Acknowledgements

This work was partially supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-272/16.12.2019.

References

Fokkens, A. S., ter Braake, S., Ockeloen, C. J., Vossen, P. T. J. M., Legêne, S., Schreiber, G., & de Boer, V. 2017. *BiographyNet: Extracting Relations between People and Events.* In Á. Z. Bernád, C. Gruber, & M. Schlögl (Eds.). *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik* Vienna, Austria: new academic press. 114–133.

Hyvönen, E., Alonen, M., Ikkala, E., & Mäkelä, E. 2014. Life Stories as Event-based Linked Data: Case Semantic National Biography. In *International Semantic Web Conference (Posters & Demos)*. pp. 1–4

Klie, J. C., Bugert, M., Boullosa, B., de Castilho, R. E., & Gurevych, I. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*. pp. 5–9

Kripke, S. 1971. Identity and Necessity. In M.K. Munitz (Ed.), *Identity and Individuation.* pp. 135–64. New York: New York University Press.

Nouvel, D., Ehrmann, M., & Rosset, S. 2016. *Named entities for computational linguistics*. ISTE.

Rospocher, M., van Erp, M., Vossen, P., Fokkens, A., Aldabe, I., Rigau, G., Soroa, A., Ploeger, T. & Bogaard, T. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37, pp. 132-151.

Tuominen, J. A., Hyvönen, E. A., & Leskinen, P. 2018. Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)* pp. 59. (CEUR Workshop proceedings; Vol. 2119). CEUR Workshop Proceedings.

Signposts for CLARIN

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

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

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

Abstract

An implementation of CMDI-based signposts and its use is presented in this paper. Arnold et al. 2020 present Signposts as a solution to challenges in long-term preservation of corpora, especially corpora that are continuously extended and subject to modification, e.g., due to legal injunctions, but also may overlap with respect to constituents, and may be subject to migrations to new data formats. We describe the contribution Signposts can make to the CLARIN infrastructure and document the design for the CMDI profile.

1 Introduction

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

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

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

Signposts are motivated with respect to the area of *growing corpora*, i.e. large corpora that are constantly extended and contain material where the conglomerate of commercial interests, intellectual property rights and privacy rights constitutes a non-trivial problem. However, all aspects signposts address are relevant to other kinds of corpora as well, generally to a different degree. In case of small and 'legally' permanent corpora, signpost information may be included in the corpus metadata. Signposts replace the concept of tombstones, which are less flexible than signposts (see Arnold et al. 2020).

2 Motivating Signposts

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

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

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

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

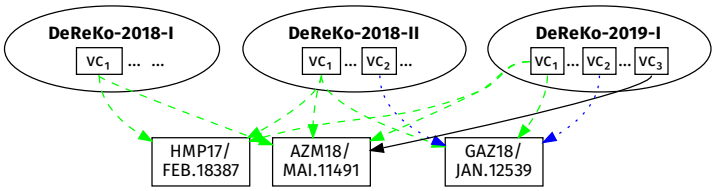


Figure 1: Relationship between DeReKo releases, virtual sub-corpora and texts (from Arnold et al. 2020). Texts may be part not of one, but many (virtual) corpora, and may belong to different versions of corpora.

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

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

In the next releases of both corpora in the IDS repository, we plan to implement signposts to avoid manually editing thousands of files in cases of conversion and legal issues.

3 A CMDI Profile for Signposts

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

The CMDI profile reuses existing components and intends to include technical information that can be automatically extracted based on a file. This information can be gained by the File Information Tool Set (FITS)[5] or other readily available tools to facilitate processing. The collected information includes the original filename, media type, file size in bytes, various checksums and cryptographic hashes. Besides this basic technical information on a LO, the signpost should also include information on the status of each LO, i.e. whether this is the currently maintained version, whether the use of the LO is deprecated (for example in the case of a migration to other data formats) or whether a file is no longer available. Reasons can be given in the provenance information (see next subsection). The schema also allows referring to a LO by a specific name that is not its filename, which is occasionally necessary.

The nature of the signpost profile is different from other profiles, in the sense that it is not intended to provide a meaningful description of a resource and does not foster findability by search engines such as the VLO. Hence

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

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

[5] See `http://fitstool.org`

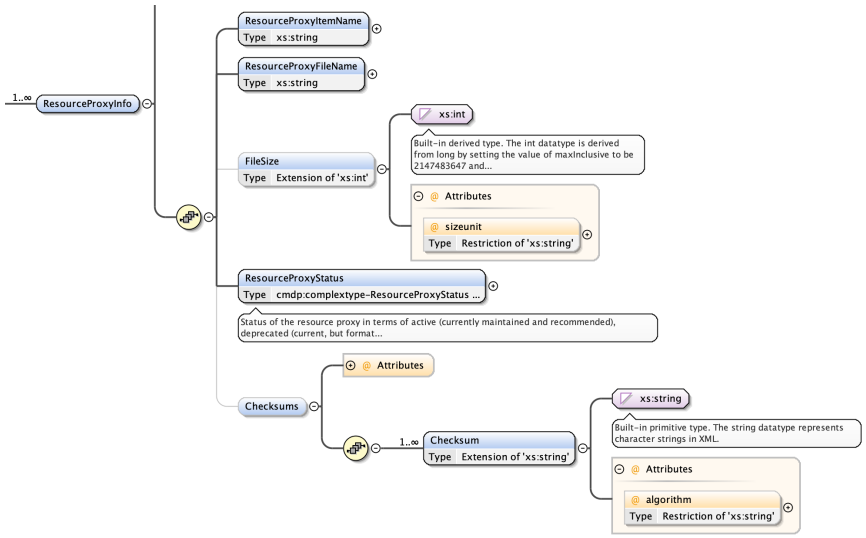


Figure 2: Visualisation of the structure for the metadata provided for each resource proxy in the CMDI Profile *signpost* (`clarin.eu:cr1:p_1587363818266`)

it does – by design – neither cater for the VLO facets, nor respect quality criteria which are automatically evaluated by the Curation Module[6]. As the media type is already provided in the resource proxy list of the CMDI file, there is no component including it. For each LO in the CMDI's resource proxy list, the profile allows the provision of various and multiple checksums, each specifying the algorithm in an attribute.

3.1 Provenance

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

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

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

[6]https://curate.acdh.oeaw.ac.at/

3.2 PIDs

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

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

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

4 Conclusion

We proposed the notion of signpost for addressing data removal, migration and deduplication in long-term archival of resources, with a specific focus on growing corpora. We also presented an implementation of the concept in the CLARIN infrastructure. We welcome feedback on the concept and on the implementation. Future work will concern the adaptation of the format, and the integrations with tools, as outlined above.

Acknowledgements

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

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

References

Arnold, Denis, Bernhard Fisseni, Paweł Kamocki, Oliver Schonefeld, Marc Kupietz, and Thomas Schmidt (2020). 'Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora'. In: *Proceedings of the LREC 2020 Workshop 'Challenges in the Management of Large Corpora' (CMLC-8)*. Marseille, France.

Marsh, Jonathan, Daniel Veillard, and Norman Walsh (Sept. 2005). *xml:id Version 1.0*. W3C Recommendation TR xml-id. The World Wide Web Consortium. URL: https://www.w3.org/TR/xml-id/.

Broeder, Daan, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel (2012). 'CMDI: a component metadata infrastructure'. In: *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*. Vol. 1.

Burnard, Lou and Syd Bauman, eds. (2020). *Guidelines for Electronic Text Encoding and Interchange. TEI P5*. version 1.0.0 2007; latest release 4.0.0 on 2020-02-13. Chicago, New York: Text Encoding Initiative.

CMDI Taskforce (2016). *Component Metadata Infrastructure (CMDI): Component Metadata Specification. version 1.2*. Tech. rep. CLARIN ERIC. URL: https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf.

De Smedt, Koenraad, Dimitris Koureas, and Peter Wittenburg (2020). 'FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units'. In: *Publications* 8.2.

ISO8879:1986 (1986). *Information processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*. Standard No. ISO 8879:1986. International Organization for Standardization.

Kupietz, Marc, Cyril Belica, Holger Keibel, and Andreas Witt (2010). 'The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research'. In: *Proceedings LREC'10*. Ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta/Paris: European Language Resources Association (ELRA), pp. 1848–1854.

Lüngen, Harald and Christopher Michael Sperberg-McQueen (2012). 'A TEI P5 Document Grammar for the IDS Text Model'. In: *Journal of the Text Encoding Initiative* 3, pp. 1–18. URL: http://jtei.revues.org/508.

Neuroth, Heike, Achim Oßwald, Regine Scheffel, Stefan Strathmann, and Mathias Jehn, eds. (2009). *nestor Handbuch. eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Version 2.0 [3/2010]. nestor.

Sperberg-McQueen, Christopher Michael and Lou Burnard, eds. (1999). *Guidelines for Electronic Text Encoding and Interchange. TEI P3*. initial release 1994-05-16; last version dated May 1999. Chicago, New York: Text Encoding Initiative.

Extending the CMDI Universe: Metadata for Bioinformatics Data

Olaf Brandt, Holger Gauza, Steve Kaminski, Mario Trojan, Thorsten Trippel
Eberhard Karls Universität
Tübingen, Germany
thorsten.trippel@uni-tuebingen.de

Abstract

CMDI is a discipline independent metadata framework, though it is currently mainly used within CLARIN and by initiatives in the humanities and social sciences. In this paper we investigate, if and how CMDI can be used in bioinformatics for metadata modelling and describing the research data.

1 Introduction

Data management in bioinformatics projects requires a very diverse and flexible set of metadata to accommodate for different scientific, organisational, and technical needs. Data categories must provide for the workflows of various types of experiments in the field of omics research (gen*omics*, prote*omics* etc.), third party suppliers such as sequencing labs, archives and public repositories. Most working groups use individual, table based metadata for their projects, which are neither semantically described, nor interoperable with established workflows in data archival or data analysis. Within the project *BioDATEN*[1] funded by the state of Baden-Württemberg in Germany, subject matter experts meet to develop an environment that facilitates data storage and collaboration of different bioinformatics working groups and archives. BioDATEN combines expertise in data management, archiving, library science, bioinformatics and related scientific workflows.

Part of ensuring the interoperability and semantic interpretation of metadata is the discussion of a common description of metadata. Though there are specific metadata schemata in the bioinformatic community like the PRIDE schema for proteomics and approaches like qPortal[2] there is no recognized gold standard for meta data handling in bioinformatics. On the other hand, there are well established standards outside bioinformatics that are used in the archiving and library community, such as METS/MODS[3], PREMIS[4], MARC 21[5], etc. The variety of research data, research questions, methods and workflows require additional flexible and research specific schemata, that can be adjusted to the needs of the concrete projects' and working groups' context. Here, the ISO standardised ISO 24622-1 and -2 and XML based CMDI framework is going to be explored as a candidate for representing the metadata in this project.

2 Motivation

Collaboration in bioinformatics is becoming increasingly important, by sharing information about genetic sequencing and the sequences for reproducing results, applying different algorithms and workflows. Sharing primary data becomes more and more widespread within the last 20+ years, but is still comparatively new to the field. This is contrasted by the fact that prices for genetic sequencing are constantly dropping, resulting in the production and availability of more data. No single data centre will be able to store and provide access to all data, even if repositories specialize for species, etc. This results in the need for a distributed infrastructure in which research data is provided in a FAIR[6] way.

Distributed environments providing data require a clear idea of the required levels of descriptions of research data. In the context of CLARIN, this has a long tradition with metadata being available and searchable with

[1] see http://www.biodaten.info
[2] see Mohr et al. 2018.
[3] see for example <METS>Version 1.6
[4] see Caplan 2009 and http://www.loc.gov/standards/premis/
[5] see https://www.loc.gov/marc/
[6] see e.g. Wilkinson et al. 2016.

tools such as the Virtual Language Observatory[7]. In bioinformatics, similar methods and tools are under developments, accompanied by strong market influences of large archives and publishers. In this paper, we try to elaborate on the technology used within CLARIN to see if the methods applied there are applicable for the BioDATEN project in bioinformatics.

2.1 Structured documentation of research data

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

In bioinformatics, we noticed that there is metadata documentation available. But there is — to our knowledge — no established and bioinformatics-wide schema. There are attempts to use `schema.org` based representations within the `bioschemas.org` project, but these are not sufficient. First, there are no profiles provided that fit OMICS data. Second, they provide a number of data category definitions, that could be utilized, but the categories extending schema.org do not have an identifier that can be used for references and it is unclear if they are stable enough for long time use. For example, the Gene Profile[8] is targeted at life sciences, including diseases, and omits processing information, while the Biosample type [9] does not provide identifiers for some data categories that could be used. However, where possible, the concepts used by bioschemas.org will be utilized here as well. The metadata can be classified according to (1) their descriptive function, (2) specific information for the community, (3) process oriented information, and (4) technical metadata. Descriptive metadata is used to describe the data in an archive for citation purposes, such as the datacite standard[10], but some data categories do not make sense in the context of bioinformatics, e.g. the concept of *Author* of raw DNA sequencing data. Process and workflow-oriented information[11] provides the background and origin of data, as well as information about the tools and experimental techniques that have been used to generate the research data[12]. Technical information contains file information often provided in terms of PREMIS. Community specific information is often provided to allow specific keywords and structures in the search process.

Existing ontologies and taxonomies are often not applied in the concrete laboratory situation, where ad hoc or laboratory specific spreadsheets are used to identify and document the research data. The lack of standard metadata formats thwarts the exchange and searchability of data with tools. The interoperability with existing portals for sharing research data thus requires tailored solutions for each lab and portal despite the fact that labs often use similar testing machinery and testing procedures. Also, publication of research data is mostly required for funding purposes and publication of scientific results. This is very similar to disciplines in the humanities such as (linguistic) annotation in fieldwork, corpus linguistics, etc.

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

Archiving research data, as well as searching and retrieving data units relies on descriptions of data units by means of metadata and the assignment of a persistent way of referencing these units. For this reason, it is essential to have a solid understanding of the data unit to be described, usually termed the *granularity* of data. Granularity in this sense is the unit of data to be stored, archived and referenced in the research process.

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

[7] see Uytvanck, Stehouwer, and Lampen 2012.
[8] see `https://bioschemas.org/profiles/Gene/0.7-RELEASE/`
[9] see `https://bioschemas.org/types/BioSample/0.1-RELEASE-2019_06_19/`
[10] see *DataCite Metadata Schema 4.3* 2019.
[11] see for example De Nies 2013.
[12] for an example of the workflow with its documentation, see Mohr et al. 2018.

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

2.3 Automatic metadata extraction requirements during a data creation workflow

Metadata creation is often seen as a burden for researchers creating data. Due to the lack of standardised processes and project management software, archiving metadata is often created manually, based for example on the headers of TEI files[13], or partly automatized by language processing applications and workflow engines such as WebLICHT[14]. The quality and completeness of the metadata in the archiving process is a major issue, and automatic metadata enrichment processes are seen as a major step forward to increase the metadata quality. This could mean to enrich metadata by authority file references, keyword extraction from textual resources, technical information extraction such as file size, checksums, dates, etc.

In bioinformatics processes, samples are analysed and processed according to complex workflows. Many of these workflows are run on HPC or cloud infrastructure, are automatized and require only little intervention, hence the manual creation of metadata is even more problematic. The creation of metadata, especially of process and technical metadata, can partly be automatized, as the workflow engines on the infrastructure use, collect and provide process information during the process. Additionally, the technical metadata can be generated easily with appropriate software tools. Larger parts of the descriptive and community specific information tend to be very similar in specialized labs, working with specific species, controlled conditions, health environments, etc. These can partly be defined in templates to be post-edited by the researcher. Again, this is similar to fieldwork situation or within large annotation projects in the humanities, though here this is often a manual process. For large NLP tasks, the metadata related processes are comparable to the bioinformatics workflows.

3 Specification and serialization options

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

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

3.1 PREMIS

Implemented and used by archives and libraries, the PREMIS standard[15] is meant to support the long-term preservation of digital objects via metadata. In the BioDATEN project, PREMIS will be primarily utilized for the storage of technical and rights metadata, as well as for the recording of events like data format conversion, checksum validation or changes in the related metadata records. The PREMIS data dictionary offers comprehensively controlled vocabularies allowing pointers with persistent identifiers. The description of scientific workflows denotes a clear limitation of the PREMIS standard. Hence PREMIS will be used for interoperability, but alone it is not sufficient for meeting all requirements.

[13] *TEI P5* 2020.
[14] see M. Hinrichs, Zastrow, and E. Hinrichs 2010.
[15] see for example Caplan 2009.

3.2 METS

In order to manage the different metadata schemas used to describe research data, it is useful to collect them in a container format. Having multiple metadata records for one digital object should be avoided. We decided to use the XML based METS format. METS is described by the an XML schema and almost exclusively serialized as XML. As a container format it is able to integrate other XML schemas without loss of information via so called extension schemas. A decisive reason to choose METS is the integration with PREMIS, which is described in detail in the literature. The different building blocks offered by the METS standard are can be used to store the variety of metadata schemas needed for research data. These schemas can be registered in METS profiles[16], which also allow for a comprehensive documentation and therefore re-usability of metadata in the METS container format. However, as a container format, METS does not provide the required metadata schemas in itself.

3.3 CMDI

Another option for modelling the metadata is by using the Component Metadata Infrastructure (CMDI, *ISO 24622-1* 2015 and *ISO 24622-2* 2019), which is also used in the CLARIN community. As an XML based serialization, many tools for editing and maintaining exist, archival system implement ways of storing the data, and transformation into other XML based formats is easy using XSLT or similar technologies. CMDI offers flexible modelling options. For example, each lab can create their own metadata profile, assembling all necessary data categories required in their respective workflow. At the same time, they can reuse parts of the metadata profiles that match the requirements of portals and service providers, archives and other partners. Using these common components, the target data format can easily be generated by a simple transformation. In fact, due to the definition of the CMDI components with concept links, for example, referring to definitions in the CLARIN Concept registry or in persistent ontologies, a high degree of semantic interoperability is achieved. For the serialization, the possible problems are similar to those already know in the CLARIN community: if labs define their own profiles and components, a certain degree of fragmentation is bound to be the result. Additionally, there is currently no fixed set of data categories, and the CLARIN Concept Registry does not contain the required definitions for bioinformatics data sets beyond interdisciplinary metadata categories. Another potential problem is that neither the bioinformatics archives and portals nor the external service providers natively support CMDI, hence a transformation is required at each step in the workflow, if CMDI were used. However, metadata generated in an automatic workflow can successively be added to the metadata file, which supports the required flexiblity.

4 Implementation

Based on previous work within CLARIN we created a CMDI Profile, the BioDaten Profile (clarin.eu:cr1:p_1588142628378, [17]. Bearing in mind the established metadata workflows of the data centres and publishers based on METS, we envision the integration of CMDI in METS-containers, also integrating PREMIS metadata. Figure 1 shows the preliminary integration of the mentioned schemes in the form of a simplified section of a METS-XML file. The complete metadata file can be viewed and downloaded from the following url: `https://bitbucket.org/steve04/cmdi-metadata-for-bioinformatics-data/src/master/METS_CMDI_PREMIS_metadata.xml`.

Using this procedure, we use the best of all worlds, using CMDIs flexibility for modelling while keeping the data interoperable with the archives and service providers.

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

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

[16] *METS Profiles* 2018.

[17] see `https://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/1.x/profiles/clarin.eu:cr1:p_1588142628378/xsd`

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

```
<mets:mets>
    ...
    <mets:metsDocumentID TYPE="UUID">6879deaa-2f64-48d7-bfc9-21cd77fb9571</mets:metsDocumentID>
    <mets:dmdSec>
        <mets:mdWrap>
            <mets:xmlData>
                <cmd:CMD>
                    <cmdp:BioDatenProfile>
                            <cmdp:GeneralInfo>
                                <cmdp:ResourceName xml:lang="en">s</cmdp:ResourceName>
                                <cmdp:ResourceTitle xml:lang="en"/>
                                <cmdp:ResourceClass>OMICS data</cmdp:ResourceClass>
                                <cmdp:Version xml:lang="en"/>
                                <cmdp:LifeCycleStatus>archived</cmdp:LifeCycleStatus>
                                <cmdp:dateCreated>2020-08-08</cmdp:dateCreated>
                                <cmdp:LegalOwner xml:lang="en"/>
                                <cmdp:FieldOfResearch>Bioinformatics</cmdp:FieldOfResearch>
                                <cmdp:Location>
                            </cmdp:GeneralInfo>
                            ...
                        </cmdp:BioDatenProfile>
                    </cmd:Components>
                </cmd:CMD>
            </mets:xmlData>
        </mets:mdWrap>
    </mets:dmdSec>
    <mets:amdSec>
        <mets:techMD>
            <mets:mdWrap MDTYPE="PREMIS:OBJECT">
                <mets:xmlData>
                    <premis:object xsi:type="premis:file">
                        <premis:objectIdentifier>
                            <premis:objectIdentifierType authorityURI="http://id.loc.gov/vocabulary/
                            identifiers" valueURI="http://id.loc.gov/vocabulary/identifiers/id">
                            identifier</premis:objectIdentifierType>
                            <premis:objectIdentifierValue>52023ca9-d9e1-40fc-bfd0-08cff7711700</premis:
```

- the Access component to provide information on accessing the resource
- the ResourceProxyListInfo component providing information on each data stream, including checksums, size, and original file name.

The tailored component SequencingInfo provides specific information on OMICS data beyond the creation process. For selecting data categories here, we were able to use Excel files used for managing metadata and provided by some partner laboratories. We also added fields that were requested by researchers.

Currently, we evaluate the mapping of laboratory internal metadata storage to this profile and assess if the integration of this metadata framework, including CMDI, METS and PREMIS in the Invenio[18] repository system, that is used within the Biodaten project. However, to our knowledge even open repository systems such as Invenio or Fedora-Commons require additional work when used with tailored metadata schemas. Additionally, it is still essential to investigate, which transformations are required from a lab internal metadata set to interoperable metadata sets used by archives and portals.

5 Future Work

At the moment it remains unclear, if and how the metadata can be enriched during automatic bioinformatics workflows. This requires APIs in the wokflow engines to extract the appropriate metadata in the process where they are present. The adaptation of the enriched metadata to specific modelling environments such as CMDI would results from this.

[18] see `https://github.com/inveniosoftware/invenio`

Acknowledgements

The work reported here was funded by the Ministry of Science, Research and Art of the Federal State of Baden-Württemberg (MWK).

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

References

Caplan, Priscilla (2009). *Understanding PREMIS*. URL: http://www.loc.gov/standards/premis/understanding-premis.pdf.

DataCite Metadata Schema 4.3 (Aug. 2019). URL: https://schema.datacite.org/.

De Nies, Tom (2013). *Constraints of the PROV Data Model*. W3C. URL: https://www.w3.org/TR/2013/REC-prov-constraints-20130430/.

Hinrichs, Marie, Thomas Zastrow, and Erhard Hinrichs (May 2010). 'WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Valletta, Malta: European Language Resources Association (ELRA).

ISO 24619 (2011). *Language resource management — Persistent identification and sustainable access (PISA)*. Standard. Geneva, CH: International Organization for Standardization.

ISO 24622-1 (2015). *Language resource management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model*. Standard. Geneva, CH: International Organization for Standardization.

ISO 24622-2 (2019). *Language resource management — Component metadata infrasctructure (CMDI) — Part 2: Component metadata specification language*. Standard. Geneva, CH: International Organization for Standardization.

<METS> Metadata Encoding and Transmission Standard: Primer and Reference Manual. Version 1.6 revised. (2010). Tech. rep. URL: http://www.loc.gov/standards/mets/METSPrimerRevised.pdf.

METS Profiles (2018). Tech. rep. URL: https://www.loc.gov/standards/mets/mets-profiles.html.

Mohr, Christopher et al. (Jan. 2018). 'qPortal: A platform for data-driven biomedical research'. In: *PLOS ONE* 13.1, pp. 1–18. DOI: 10.1371/journal.pone.0191603.

Guide to generate PRIDE XML files (n.d.). URL: https://www.ebi.ac.uk/pride/help/archive/submission/pridexml.

TEI P5 (2020). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. URL: https://tei-c.org/Vault/P5/4.0.0/doc/tei-p5-doc/en/html/index.html.

Uytvanck, Dieter van, Herman Stehouwer, and Lari Lampen (May 2012). 'Semantic metadata mapping in practice: the Virtual Language Observatory'. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 1029–1034.

Wilkinson, Mark D. et al. (2016). 'The FAIR Guiding Principles for scientific data management and stewardship'. In: *Scientific Data* 3.160018. URL: https://doi.org/10.1038/sdata.2016.18.

The CMDI Explorer

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

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

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

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

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

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

Abstract

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

1 Motivation

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

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

2 Background

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

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

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

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

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

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

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

Other tools face similar deficits. The CLARIN community offers a number of converters from CMDI to bibliographic metadata standards such as Dublin Core or MARC 21 (Zinn et al., 2016). These converters

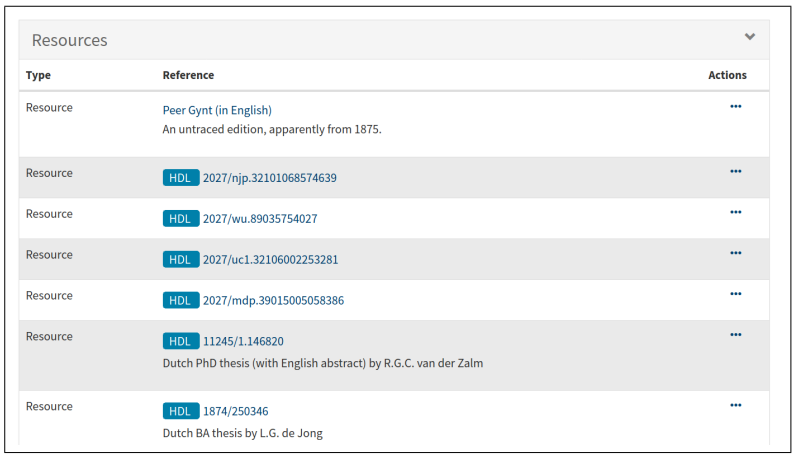


Figure 2: A simple structured CMDI record in the VCR.

either show similar shortcomings when it comes to processing highly structured CMDI files, or are tailored to specific CMDI profiles where hence structural complexities are known in advance.

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

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

3 The CMDI Explorer

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

The CMDI Explorer is designed to analyse highly recursive CMDI files and to provide a tree-based representation and visualisation of its entire structure, both in the browser and as a downloadable HTML and JSON file. The CMDI Explorer can also work with resources composed of constituents that are described using different CMDI profiles: As it only relies on the resource proxy list, it will just assemble all files referenced there irrespectively of their role in the CMDI file. As it is designed to work with collections, the CMDI Explorer can operate across repositories, provided the metadata is freely available,

i.e., not behind an Authentication and Authorization Infrastructure (AAI) wall.

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

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

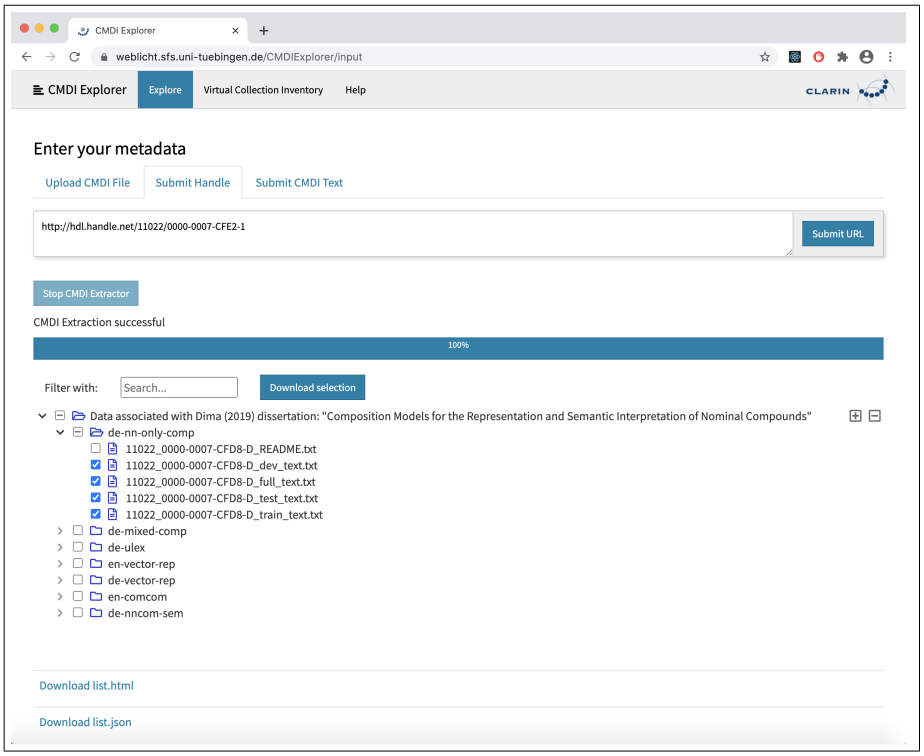


Figure 3: A CMDI record displayed in the CMDI Explorer

There were a number of challenges involved with the extraction of the data. Firstly, it is not always straightforward to extract the CMDI XML data associated with a PID. For each PID URL, there are a number of redirects to go through before arriving at the final URL. The path that the redirects take is also affected by the value of the `Accept` header. It was found that this needed to be set to `application/x-cmdi+xml, application/xml+cmdi, application/xml` for both following the redirects as well as extracting the data from the URL to cover the different possible Media Types the CMDI XML data could have as its `Content-Type`. There are also a number of CMDI metadata resources in certain collections that are mislabelled as Resource as their resource type, when the resource type should be Metadata. In order to solve this, all resources labelled as Resource with an XML, CMDI, or no Media Type are analyzed and checked if they can be analyzed as a CMDI file, i.e., checking whether the file contains a resource proxy list which is non-empty. If so, then the resource is analyzed as CMDI metadata. Another problem is that some resources are inaccessible for various reasons, such as the need for authentication. Inaccessibility of resources is indicated in the data tree and the corresponding HTML or JSON files. Moreover, some collections are 'circular', *i.e.*, a resource in the CMDI metadata of one resource will eventually lead back to the original resource. The Explorer keeps

records of resources encountered so far, and any resource that has already been analyzed is simply not included in the tree that is being constructed.

Front-end. Fig. 3 shows the main interface, presenting users the three options to enter their input.

Once the metadata has been submitted, the Explorer's back-end attempts to extract its underlying tree. During extraction, users see a progress bar, and live updates to the emerging tree visualisation. Fig. 3 shows a top node together with its immediate children. Each branch of the tree can be unfolded individually, and there are also two controls ($+$, $-$) to unfold or fold the entire tree. Leaf nodes have actions attached to them: (i) copy the handle to the clipboard, (ii) send the handle to the Switchboard, (iii) click on the handle so that it resolves in the browser, and (iv) download the resource referred to by the handle. Also, the entire tree can be downloaded in a structure-preserving HTML or JSON format. Moreover, users can select nodes individually. Selected nodes will be added to the 'download basket'; the corresponding data resources of their selection are made available as a ZIP archive file.

4 Current State and Future Work

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

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

The CMDI Explorer has already been included in the test version of the Switchboard as a metadata processing tool. We expect the Explorer to be connected to the VLO and the VCR as well so that users can easily invoke it wherever they find complex CMDI metadata they need to explore further.

Acknowledgements

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

References

Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, volume 1.

Willem Elbers. 2017. Virtual collection registry v2. Technical report, CLARIN ERIC.

Margaret King, Davor Ostojic, Matej Ďurčo, and Go Sugimoto. 2016. Variability of the facet values in the vlo – a case for metadata curation. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015*, pages 25–44, Linköping. Linköping University Electronic Press.

Davor Ostojic, Go Sugimoto, and Matej Ďurčo. 2017. Curation module in action – preliminary findings on vlo metadata quality. In *Proceedings – CLARIN Annual Conference 2016*.

Dieter van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari et al., editor, *Proceedings of LREC'12*, Istanbul, Turkey. ELRA.

Claus Zinn, Thorsten Trippel, Steve Kaminski, and Emanuel Dima. 2016. Crosswalking from cmdi to dublin core and marc 21. In Nicoletta Calzolari et al., editor, *Proceedings of LREC'16*, Portorož/Paris. ELRA.

Claus Zinn. 2018. The language resource switchboard. *Comput. Linguist.*, 44(4):631–639.

Matej Ďurčo. 2013. *SMC4LRT – semantic mapping component for language resources and technology*. Ph.D. thesis, Technische Universität Wien.

Going to the ALPS: A Tool to Support Researchers and Help Legality Awareness Building

Veronika Gründhammer
Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)
Austrian Academy of Sciences
veronika.gruendhammer@oeaw.ac.at

Vanessa Hannesschläger
Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)
Austrian Academy of Sciences
vanessa.hannesschlaeger@oeaw.ac.at

Martina Trognitz
Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)
Austrian Academy of Sciences
martina.trognitz@oeaw.ac.at

Abstract

In this paper, we describe the "ACDH-CH Legal issues Project Survey" (ALPS), a tool that helps researchers to understand all legal dimensions of their research project and the lifecycle of their data. The introduction explains the institutional preconditions and specific target groups of the tool, which was developed by the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH) of the Austrian Academy of Sciences and is used to support both its members and project partners to comply with legal requirements and the spirit of Open Science at the same time. The paper then focuses on the various elements of the survey and their goals, also explaining the workflow that is used to process the results of survey participants. The conclusion and outlook section suggests ways to open up this tool for use by a community beyond the members and project partners of the hosting institution.

1 Introduction: Institutional preconditions

At the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH, formerly Austrian Centre for Digital Humanities ACDH), Open Access to publications and research data as well as Open Source regarding software have always played an important role. Furthermore, the Austrian Academy of Sciences has been promoting Open Access to research outputs through funding schemes such as *go!digital*, which aim to improve the framework conditions for data-based and data-driven research in the humanities. This also means that the data that should be the basis for the research need to be openly available.

In February 2019, the ACDH released its Guidelines on Managing Research Data, Results, and Software (in short: "Open Policy", original title "Richtlinien für das Management von Forschungsdaten, -ergebnissen und Software am ACDH-OeAW"; cf. Hannesschläger 2019). The publication of this Policy[1] - which is based on a template developed in the project *e-Infrastructures Austria* (2016) - is therefore a natural consequence of the institute's, but also the Academy's commitment to Open Science.

[1] Currently, Austria has no national policies concerning Open Access, Open Research Data and Open Science. However, according to ROARMAP, 14 Austrian research organisations and two funders already have one (cf. http://roarmap.eprints.org/cgi/search/advanced).

When following an Open Science approach and making publications and/or data openly accessible it is crucial to consider a number of legal issues, which researchers are often not familiar with (cf. Kamocki et al. 2016; 2018). Examples include the following:

- Licensing as one of the key topics in research data management
- Processing of personal data in compliance with the principles and requirements laid down by the General Data Protection Regulation (GDPR)
- Disclosure of certain information in the form of an imprint according to §25 of the Austrian media law / e-commerce law

As researchers are not always aware of all of these areas and how they concern their work, the ACDH-CH has been striving to take measures to educate its members and project partners. Expertise and teaching material has been gathered and developed at the institute, a number of lectures and workshops have been organized and active involvement in the DARIAH working group "Ethics and Legality in Digital Arts and Humanities" (ELDAH) and the CLARIN Legal and Ethical Issues Committee (CLIC) as well as the Legal Issues Working Group of the Open Science Network Austria (OANA) is being pursued. Basic information and a general introduction to the most common DH-relevant aspects of the law have also been published in the "Services" section of the ACDH-CH website ([2019]). In order to further these efforts and to collect information about possible legal problems from ACDH-CH researchers and projects in a structured manner, the ACDH-CH Legal issues Project Survey (ALPS) was developed. The survey to be presented in more detail in the following section has to be seen in relation to the above mentioned aspects.

2 Goals and target groups

The goal of ALPS is to provide a standardized and comprehensive overview of legal areas that digital humanities research is concerned by or with. On the one hand, the survey aims to collect information about the legal situation of all projects carried out at or in cooperation with the ACDH-CH in order to ensure that no fundamental legal misconduct is made by ACDH-CH researchers or projects. On the other hand, ALPS shall support its users, i.e. the researchers answering the questions. The survey should help them understand what legal areas concern them and how to comply with national and international legal requirements. It is an awareness building tool and therefore also provides links to further resources and reading so that the researchers have the possibility to learn more about the laws that concern them. Project leads are the primary target group of the tool, but also all other researchers and developers. Currently, ALPS is an institute-internal tool and therefore only targets these user groups as ACDH-CH members. However, opening up the tool for general use is planned (see outlook).

3 Development and implementation

ALPS was developed in a dedicated taskforce for legal issues at the ACDH-CH. The group consists of six members with experience in general project administration, Austrian copyright and data privacy legislation, Open Science, software and other types of (open) licensing, and digital archiving. This group was established at the end of 2017 and is responsible for supporting research projects carried out by or in cooperation with the ACDH-CH with advice on legal aspects of digital research.

Due to the numerous requirements for projects mentioned in the section "Institutional preconditions" above, a standardized questionnaire was developed to help in gathering information about the status quo of ongoing projects. This questionnaire first took the form of a fill-in google document, a separate copy of which was created for each individual project. Interviews were then conducted by taskforce members in face to face meetings with project representatives. Any feedback from their side as well as any clarification needs were fed back into the questionnaire to improve its understandability and usability. After the first interviews, a larger internal workshop with about 20 participants was carried out at the ACDH-CH to discuss and improve the questionnaire's practicality.

After the questionnaire content had thus been sufficiently developed and fine tuned, it was transformed from a google document into a proper online survey using the software LimeSurvey (cf. https://www.limesurvey.org/) by the authors of this paper. LimeSurvey is a free and Open Source software for implementing online surveys for a large variety of needs and disciplines. It also allows for extensive statistical analysis as well as custom export of survey answers for further analysis.

4 The survey and its individual parts

ALPS consist of 72 questions in total. They are split up into several sections that contain various types of questions on specific legal topics. ALPS participants answer questions about the imprint on the project website, about the provenance and legal situation of reused data, about the Openness of data and code created in the project, and about data privacy issues.

The first section collects general information about the project such as the title, the PI, and information about cooperation partners (who might be responsible for certain legally relevant areas, e.g. data creation or personal data processing).

Section two deals with imprints, which according to §25 of the media law or e-commerce law (for websites) need to disclose certain information about the media owners themselves, information about the editors as well as information on the basic thematic direction of the website.

The third questionnaire section concerns all projects that use existing software and/or produce new code. Questions about the provenance and licenses of reused code, about the rights attached to newly created code, and about the proper documentation of these rights are asked. References to useful resources such as the CLARIN LINDAT Public License Selector (cf. https://ufal.github.io/public-license-selector/), which enables the users to check the compatibility of code licenses, are provided.

In the fourth part of the questionnaire, projects that produce new data or databases are the topic of interest. It focuses on questions in the area of licensing and opening up data. Among the resources provided for users in this section are the ACDH-CH guidelines on "Legal aspects of Digital Humanities projects" ([2019]).

In the subsequent section, all projects working with existing data or databases are asked to give answers. The main focus is the legal re-usability and the possibility of openly sharing data.

The survey section on data privacy issues is split up in two parts. In a first part, general questions about data privacy are asked with the aim to collect information and to build awareness. The second part of the data privacy section asks questions specifically tailored to collect information about possibly created data applications. As the taskforce has to collect and pass on information about all data applications run by the ACDH-CH to the Academy's data protection officer, this section of the survey is designed to collect information in the form in which it has to be inputted in the Academy's data application management software (NioBase, cf. https://niobase.com/).

In the final section, survey participants can inform the taskforce if they feel that they discovered problematic legal aspects while filling in the survey. If so, they have the opportunity to request a follow-up face to face meeting and specify what they would like to talk about.

5 General workflow

The general workflow for obtaining information about the legal status of research projects carried out at or in cooperation with the ACDH-CH is designed as follows: When a new project begins, the ACDH-CH projects manager (who has an overview of all projects starting anew) invites the principal investigator or project lead (PI) to participate in the survey by sharing a link to ALPS. This happens only after a project has been approved and initiated, as many of the questions to be answered might not yet be clarified at proposal stage. The PI can subsequently fill in the survey and has the option to pause, save and resume filling in the survey at a later point in time. This also enables PIs to only fill in parts of the survey and then pass on the link to other project members who might be more familiar with a certain aspect of the project (e.g. the developer will likely be more familiar with the specifics of software licenses relevant for the project than the PI). If the survey is not completed within three weeks, bi-weekly email reminders are sent to the PI. When the survey is completed, the task force members receive a notification from the LimeSurvey environment. At least two task force members check the replies and decide jointly if any follow up action is necessary, either by request of the project team or due to issues arising from the provided answers.

6 Conclusion and outlook

Currently, ALPS is only available in English because this ensures a maximum coverage of ACDH-CH institute members, who all work in a German speaking country but are not all native German speakers. However, one of our next goals is to translate the survey to German in order to provide a language choice to researchers who feel more comfortable answering delicate questions on legal issues in their mother tongue. LimeSurvey already comes with built-in support for multiple languages.

Thanks to the technical LimeSurvey infrastructure, this way of asking researchers about the legal preconditions of their work will allow us to collect long term statistics about the most prevalent and pressuring legal issues and the in- or decrease in awareness about legal issues among ACDH-CH researchers. The re-evaluation of the survey structure and answers will be an ongoing process as the law is not static and therefore, neither is ALPS.

As described in this paper, ALPS currently is an internal evaluation and awareness building tool to make sure that all ACDH-CH projects fulfill legal and quality standards. However, its current internal use already shows the great potential that this survey tool can have for supporting members of a larger infrastructure such as CLARIN or DARIAH. Therefore, we hope to take up discussions with CLARIN, specifically with members of the CLARIN Legal and Ethical Issues Committee CLIC, about implementing this tool as a service for the CLARIN community. In a first step, it might be adapted for use within CLARIN Austria, as the questionnaire has been designed to match Austrian legislation. Subsequently, the questionnaire could be taken to a more generalized level in order to fit other national legislations in CLARIN member countries, respectively become legislation-agnostic on the national level and only take legislation on the EU level into account. One of the central issues to discuss in this context is the fact that while ALPS is a digital tool, the long-term support for its users requires substantial personnel resources, as experts on legal questions have to be available to users if questions arise (and they always do). Active working groups within the ERICs such as the CLIC are therefore vital for the sustainability of such a tool in the context of research infrastructures.

References

ALPS [ACDH-CH Legal issues Project Survey]. https://survey.acdh.oeaw.ac.at/index.php/414341

ACDH-OeAW. 2019. Richtlinien für das Management von Forschungsdaten, -ergebnissen und Software am ACDH-OeAW [Guidelines on Managing Research Data, Results, and Software at ACDH-OeAW].

ACDH-CH. [2019]. Legal aspects of Digital Humanities projects. https://www.oeaw.ac.at/acdh/services/legal-aspects-of-dh-projects/

Austrian Academy of Sciences. go!digital. https://www.oeaw.ac.at/en/funding/subsites/godigital/

e-Infrastructures Austria. 2016. Model policy for research data management (RDM) at Austrian research institutions. v1.2. https://services.phaidra.univie.ac.at/api/object/o:459162/diss/Content/get

Vanessa Hannesschläger. 2019. Nachhaltigkeit durch Institutionalisierung: Die Open Policy des Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften. Leibniz University Hannover, Open Access Tage 2019. Zenodo. http://doi.org/10.5281/zenodo.3465505

Paweł Kamocki, Erik Ketzan, and Julia Wildgans. 2018. Language resources and research under the General Data Protection Regulation. CLARIN Legal Issues Committee CLIC White Paper Series, CLIC White Paper #3. https://www.clarin.eu/sites/default/files/CLIC_White_Paper_3.pdf

Paweł Kamocki, Pavel Stranák, and Michal Sedlák. 2016. The Public License Selector: Making Open Licensing Easier. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, ed. by Nicoletta Calzolari et al., 2533–2538. Paris: European Language Resources Association (ELRA), 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/880_Paper.pdf

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 / GDPR). http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

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

Paweł Kamocki
Leibniz-Institut für Deutsche Sprache
Mannheim, Germany
kamocki@ids-mannheim.de

Abstract

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

1. Introduction

N-grams, sequences of n-items from a sample of text, are of utmost importance for modern linguistics and language technology. For example, the use of Google N-gram Viewer has become commonplace in the language community, despite its questionable quality and lack of metadata (Koplenig, 2017). It is therefore not surprising that linguists attempt to compile their own re-usable lists of n-grams. To this end, one of the most recurrent questions that language researchers ask legal experts is: can n-grams be freely used and shared? If so, how long should they be?

The only possible *in abstracto* answer to this question is quite disappointing: while in general short n-grams can be used and shared without consequences (at least from the copyright perspective), there are no clear limits as to the length of copyright-free n-grams. The decision on where to draw the line should be made for every project on a case-by-case basis. This paper will hopefully provide some guidance on this issue.

2. The Traditional Approach, or 'Originality, You Fool!'

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

Following this approach, the Court of Justice of the European Union ruled in the *Infopaq* case (probably the most important case concerning text snippets): *'it should be borne in mind that there is nothing in [EU directives] indicating that [parts of works] are to be treated any differently from the work as a whole. It follows that they are protected by copyright since, as such, they share the originality of the whole work. (...) the various parts of a work thus enjoy protection under [the Infosoc Directive], provided that they contain elements which are the expression of the intellectual creation of the author of the work'.*[1]

This very same judgement was also the Court's first attempt to harmonise the notion of 'originality' across EU Member States (Rosati, 2013) as 'the author's own intellectual creation' (which exists in some EU Directives,[2] but also, accidentally, is the traditional definition of originality in German copy-

[1] CJEU, 16 July 2009, C-5/08 (*Infopaq*)

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

right: *persönliche geistige Schöpfung*). The Court further elaborated on this definition in the *Painer* case,[3] where it ruled that a work is an intellectual creation of the author if it reflects his personality and expresses his free and creative choices.

The possibility to exercise choice in the creative process is therefore a necessary (if not sufficient) condition of originality. One could then hypothesise that the choice of a single word can be sufficient to meet this requirement. The 2008 French Court of Cassation decision[4] declaring a work consisting of a single word (*Paradis*) original could be quoted to support this statement. However, it should not be forgotten that the word was not only written in a very specific font, but also - more importantly - placed in a very specific place: above the toilet of a mental hospital. In other words, the protection was not granted to the word *'Paradis'* as a literary work, but rather to the whole setting, which constituted an artistic (and not literary) work.

Another international definition of originality - formally concerning compilations, but which can be applied more generally (after all, literary works for linguists are but compilations of words) - states that it should manifest itself in the 'selection and arrangement' of various elements constituting a work. This criterion - which, interestingly, stems from a canonical US copyright case *Sarony* concerning a lithography of Oscar Wilde[5] - appears not only in the Berne Convention (Article 2.5), but also, in a slightly 'updated' form, as an alternative (selection OR arrangement), in the TRIPS Agreement (Article 10.2), the EU Database Directive (Article 3.1) and numerous national laws (e.g. Article L112-3 of the French Intellectual Property Code or Section 4 of the German Copyright Act). In its original form (as a conjunction), known from the Berne Convention or from the *Sarony* case, this criterion would imply that mere choice (selection) is not enough to constitute originality, and that another aspect: arrangement is also necessary. In other words, the constitutive elements not only have to be chosen by the author, but also placed in a particular order, which in the context of snippets would imply that two words is the absolute minimum for a snippet to be original.

It is indeed our opinion that in 'pure' international copyright two words can be enough - still, only in extremely limited cases - to constitute an original work, or an original snippet. The position according to which two-words snippets can (in very rare cases) be original is also supported by Article 2.1 of the Berne Convention, according to which copyright protection is independent from 'the mode or form' in which the work is expressed, which arguably includes also very short forms. This same rule can also be found in Article L112-1 of the French Intellectual Property Code.

This seems to be the position of French judges, who found such two-words combinations as e.g.: *du rififi*[6], *Charlie Hebdo*[7], *Bourreau d'enfants*[8] or *Paris Canaille*[9] to be protected by copyright as original titles.[10] These decisions, however, are rather old. German courts seem to be more demanding when it comes to originality of very short works. The shortest work declared copyright-protected by a German court that we have been able to identify is a four-words slogan *Ein Himmelbett im Handgepäck*[11] (a canopy bed in hand luggage); however, the decision dates back to 1964. Modern German case law seems to generally refuse copyright protection of slogans and titles, which are deemed too short to constitute original works. This is also the position in the United States, where the Copyright Office

---

3 CJEU, 1 December 2011, C-145/10 (*Painer*)

4 Cour de cassation, 1 Civ., 13 novembre 2008, no. 06-19.021

5 U.S. Supreme Court, 17 March 1884, *Burrow-Giles Lithographic Company v. Sarony*, 111 U.S. 53

6 Cour d'appel de Paris, 4e ch., 24 janvier 1970, RTD com. 1971, p. 94, obs. H. Desbois

7 Cour d'appel de Paris, 4e ch., 25 octobre 1995, JurisData n° 1995-024506

8 Tribunal de Grande Instance de Seine, 3e ch., 2 février 1960, RTD com. 1960, p. 844, obs. H. Desbois

9 Cour d'appel de Paris, 1er ch., 30 mai 1956, *Léo Ferré c/ Sté Océan Films et a.*, JCP G 1956, II, 9354

10 Article L112-4 of the French Intellectual Property Code states that 'Le titre d'une oeuvre de l'esprit, dès lors qu'il présente un caractère original, est protégé comme l'oeuvre elle-même'; however, it can be argued that titles are protected by *sui generis* copyright, which only restricts the use of an original title as a title of another work.

11 Oberlandegericht Düsseldorf, 28 Februar 1964 – 2 U 76/63.

states that *'short phrases, such as names, titles, and slogans, are uncopyrightable because they contain an insufficient amount of authorship'*.[12]

So when does a literary work become 'long enough' to be considered for copyright protection according to modern standards? As stated above, there is no definitive answer, but some guidance has been provided by the CJEU in the above-mentioned *Infopaq* case. The Court ruled that snippets of 11 consecutive words can be original (although the evaluation of their actual originality was, of course, left to the national court). The 11-words *limes* resulted simply from the facts of the case (this was the length of snippets used by *Infopaq*, an early news aggregator service), and the case should not be interpreted as meaning that 10-words or shorter snippets are free from copyright; it does, however, provide an argument in the discussion. The 10-words limit for snippets should, in our opinion, be considered as very liberal. The truth lies therefore somewhere between 2 and 10.

3. The New Approach, or 'Name that Tune'

A different approach, not requiring originality of the used part of a work, was adopted by the CJEU in a recent case *Pelham*.[13] The facts involved not a literary work, but a sound recording (phonogram) by Kraftwerk, a short part of which ('approximately two seconds rhythm sequence') was used as a sample in another recording. Rather than considering the sample's originality (which, arguably, is very difficult to evaluate out of context), the CJEU ruled that its use constitutes 'reproduction in part' (and therefore an act that in principle requires authorisation of the rightholder) if it is 'recognisable to the ear'.

How to apply this approach to text snippets? Arguably, it means that n-grams can be reused as long as they are not *hapax legomena* (i.e. as long as they occurred independently in more than one text in the language, to the extent that this can be established), and as such their exact source cannot be identified. In this approach, the longer the snippet, the more likely it is to be unique, i.e. to be a *hapax legomenon*. This test seems to be generally more strict than 'originality' (e.g. a purely descriptive, banal paragraph will at some point become a *hapax legomenon* if it is long enough, but it will still lack originality), but it has the advantage of being more objective, and therefore perhaps easier to apply *in abstracto*.

This approach presents a theoretical weakness: it can be argued that it was not intended to apply to copyright, but only to the related right of phonogram producers, as the CJEU ruling was specifically about the interpretation of Article 2(c) of the InfoSoc Directive 2001/29/EC (concerning reproduction of phonograms), and not the whole Article 2 (concerning reproduction in general). Indeed, very short sequences from sound recordings seem to be easier to recognise than very short snippets of text, so the 'recognisability' test may render more reliable results when applied to phonograms than to literary works.

However, this approach also presents a practical advantage, as it is quite commonsensical: if the excerpt is not recognisable to the rightholder, then he will not sue for copyright infringement, and if it is not recognisable to the judge, its use will not be qualified as copyright infringement.

4. New Related Right of Press Editors: A Trench War Has Begun

The new Directive 2019/790 on Copyright in the Digital Single Market (DSM Directive) introduced and harmonised a new related right of press editors. The right protects said editors against 'parasitic' use of press articles by commercial online services, and as such is not directly relevant for language research (Papadopoulou, Moustaka, 2020). However, this new right is only triggered when an online service uses more than 'individual words or very short extracts of a press publication'. It may therefore be necessary to define precisely, either via case law or via a collective agreement, the amount of consecutive works that can be freely used. In Germany, where this right was first introduced in 2013 (before the DSM Directive), the Patent and Trade Mark Office initially (2015) recommended 7 consecutive words as a freely reusable 'very short extract' of a press publication. However, this recommendation no longer appears on its website.[14]

---

[12] US Copyright Office, Circular 33: Works Not Protected by Copyright.

[13] CJEU, 29 July 2019, C-476/17 (*Pelham*)

[14] But cf. the 2015 news report at: https://www.internet-law.de/2015/09/leistungsschutzrecht-dpma-schlaegt-einigung-vor.html (retrieved 4 September 2020).

The debate on this subject has recently been revived in Germany by the publication of a government bill on the implementation of the DSM Directive.[15] The bill does not define the precise length of a 'very short extract', but instead states that headlines (*Überschriften*), which can sometimes be quite long (even longer than 7 words), should be considered as such. In a joint statement, three German associations of press publishers (BDZV, VDZ and VDL) have starkly disapproved this part of the bill, arguing that 'very short extracts' should be limited to three consecutive words (BDZV et al., 2020). This may be an indication that, also in the context of other uses, press publishers will not oppose the use of 3-grams extracted from their articles. In the near future, the German legislator may adopt a more precise position on the question (likely defining the 'very short extract' as containing between three and seven consecutive words), which could then, in our opinion, also be extrapolated to other contexts.

5. Conclusion

As promised in the introduction, we will try to formulate some guidelines concerning the use of short snippets of text without permission from the rightholders, in cases not covered by statutory exceptions (e.g. the exception for non-commercial scientific research):

- 3-grams should be regarded as generally free from copyright, even under the most conservative of the presented approaches;
- 7-grams is a reasonable consensus proposed by the German Trademark and Patent Office, albeit in a specific context other than copyright;
- a very liberal interpretation of the CJEU's *Infopaq* ruling may lead to the conclusion that 10-grams are free from copyright; however, we feel that this length is excessive;
- those who are pragmatically-minded may be tempted by the 'recognisability theory', under which a part of a work may be freely re-used as long as the work that it was taken from cannot be identified. Under this theory, it seems that even relatively long snippets (such as 10 or more words) of prose can possibly be used, especially if some 'strongly identifying' elements such as titles or names are removed.

The guidelines presented above only apply if the use of the data is not based on a license (in which case, of course, the license takes precedence), or covered by a statutory exception.

References

BDZV (Bundesverband Deutscher Zeitungsverleger Verband), Deutscher Zeitschriftenverleger Verband, Deutscher Lokalzeitungen. 2020. *Stellungnahme zum Diskussionsentwurf des Bundesministeriums der Justiz und für Verbraucherschutz für ein Erstes Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts vom 16. Januar 2020*.

Koplenig, Alexander. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169-188.

Papadopoulou, Maria-Daphne and Evanthia-Maria Moustaka. 2020. Copyright and the Press Publishers Right on the Internet: Evolutions and Perspectives. In: Tatiana-Eleni SynodinouPhilippe JougleuxChristiana Markou-Thalia Prastitou [Eds.]. *EU Internet Law in the Digital Era*. Springer: Cham.

Rosati, Eleonora. 2013. *Originality in EU Copyright: Full Harmonization Through Case Law*. Edward Elgar Publishing: Cheltenham, Northampton.

[15] *Diskussionsentwurf des Bundesministeriums der Justiz und für Verbraucherschutz Entwurf eines Ersten Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts, 15 January 2020*, available at: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/DiskE_Anpassung%20Urheberrecht_digitaler_Binnenmarkt.pdf?__blob=publicationFile&v=1 (retrieved 4 September 2020).

CLARIN contractual framework for sharing language data: the perspective of personal data protection

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

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

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

Pawel Kamocki
IDS Mannheim, Germany
pawel.kamocki@gmail.com

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

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

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

Mari Keskküla
University of Tartu, Estonia
mari.keskkula@gmail.com

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

Abstract

The article analyses the responsibility for ensuring compliance with the General Data Protection Regulation (GDPR) in research settings. As a general rule, organisations are considered the data controller (responsible party for the GDPR compliance). Research constitutes a unique setting influenced by academic freedom. This raises the question of whether academics could be considered the controller as well. However, there are some court cases and policy documents on this issue. It is not settled yet.

The analysis serves a preliminary analytical background for redesigning CLARIN contractual framework for sharing data.

1 Introduction

This paper focuses on sharing[1] language data (LD) containing personal data (PD).[2] A key issue here is to define obliged parties under the General Data Protection Regulation (GDPR). The GDPR identifies

[1] The General Data Protection Regulation (GDPR) Art. 4 (2) defines processing extensively so that it covers any operation which is performed on personal data (e.g., collection, storage, alteration and sharing).

[2] The article does not address transfer of PD outside the EU. This topic concerns special provisions of the GDPR and EU case law (e.g. C‑311/18).

the controller and the processor as responsible parties (Art. 4).[3] According to the accountability principle, the controller is responsible for the GDPR compliance (Art. 5 (2)). The GDPR defines the controller as "*the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data*" (Art. 4 (7)). The processor is defined as "*a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller*".[4]

The authors place this question into the context of language research and address practical questions such as 1) how to differentiate between obligations of the university and individual researchers (e.g., whether an individual researcher is the controller) and; 2) whether sharing LD results in joint controllership or separate controllership (whether the transferee of the data becomes the controller, the joint controller or the processor). The paper serves as a preliminary conceptual analysis behind redrafting CLARIN contractual framework for data sharing. It develops further the previous research (see Kelli et al. 2015; Kelli et al. 2018) focusing on personal data.

2 Duty-bearers for the GDPR compliance within research settings

The key feature of the controller is the **determination** of the **purposes** and **means** of the processing of PD.[5]

A relevant issue for the research community is to analyse whether an individual researcher and/or the university is the controller. Research settings is a unique context since academic freedom is defined as a fundamental right.[6] This gives researchers often more freedom compared with other employees and complicates matters as seen from the analysis.

Considering relations between companies and their employees WP29 (2010: 15) indicates that "preference should be given to consider as a controller the company or body as such rather than a specific person within the company or the body. It is the company or the body which shall be considered ultimately responsible for data processing and the obligations stemming from data protection legislation". According to the European Commission (EC) "if your company/organisation decides 'why' and 'how' the personal data should be processed, it is the data controller. Employees processing personal data within your organisation do so to fulfil your tasks as data controller". DP Handbook (2018: 102) similarly asserts that legal entity (not employees) is the controller.

WP29 (2010: 6) further explains that "*Controller and processor and their staff are therefore considered as the 'inner circle of data processing' and are not covered by special provisions on third parties*"". This is compatible with employment law. The employment law literature and case law also set forth that one of the characteristics of an employee is the fact that the employee is merged with the employer's team of other employees. The employee is employed and acts within the framework of the employer's economic activities (Risak & Dullinger 2018; C-22/98 para 26; C-66/85; C-256/01).

Although individual researchers conduct research with some freedom, their work is coordinated by the university. The university is responsible for the GDPR violations[7], and it has to implement appropriate technical and organisational measures (incl. data protection policies) to ensure the protection of PD (GDPR Art. 24) and to maintain a record of processing activities (Art. 30 GDPR).[8] For instance, to

[3] When it comes to liability, then as a general rule, any person who has suffered damage as a result of the GDPR violation has the right to receive compensation from the controller or processor (GDPR Art. 82 (1)).

[4] WP29 (2010: 1) explains that the processor has to meet two conditions: 1) a separate legal entity with respect to the controller; 2) it processes PD on behalf of the controller.

[5] According to WP29 "*Determination of the "means" includes both technical and organizational questions where the decision can be well delegated to processors (as e.g. "which hardware or software shall be used?") and essential elements which are traditionally and inherently reserved to the determination of the controller, such as "which data shall be processed?", "for how long shall they be processed?", "who shall have access to them?", and so on'*"" (2010: 14).

[6] According to the EU Charter of Fundamental Rights "*The arts and scientific research shall be free of constraint. Academic freedom shall be respected*" (Art. 13).

[7] Pursuant to WP 29 (2010: 15) "*In the strategic perspective of allocating responsibilities, and in order to provide data subjects with a more stable and reliable reference entity for the exercise of their rights under the Directive, preference should be given to consider as controller the company or body as such rather than a specific person within the company or the body. It is the company or the body which shall be considered ultimately responsible for data processing and the obligations stemming from data protection legislation*".

[8] There might be a practical problem when an individual researcher collects data himself and his host university is unwilling to be the controller. One approach could be that if the host university of a researcher is unwilling to become the controller,

ensure the GDPR compliance the University of Tartu (Estonia) has adopted the Data Protection Policy, which implies its position as the controller. In case academics (researchers, professors) infringe personal data rights, then they are liable before the university. The question is whether academics could be considered the controller as well. The issue is not settled yet. However, European Parliamentary Research Service in its study concerning research (EPRS study 2019: 34) suggests that "*researchers and universities should assume that when processing personal data, their activities render them data controllers*".

The analysis is likely to be different in countries (like Germany or France) where academics (or at least university professors) are not employees, but public servants appointed for life and independent in the exercise of their missions (not unlike, e.g. judges). It is due to this independence that in both Germany and France, e.g. copyright in the works created by academics belongs to them, and not to their university or institution.[9] In light of this rule (referred to as '*professors' privilege*'), it may be hard to argue that professors reap profits of their work (because they are free to decide how to do it), but the university should bear the responsibility for how they process personal data. Along these lines, the French National Centre for Academic Research (CNRS) in its guide on data processing for research purposes (CNRS guide 2019: 12) defines the director of a unit (so an individual, not an institution) as the data controller. Then, the director of a unit designates the CNRS' Data Protection Officer as DPO, thereby providing for a sort of 'bottom-up centralisation' of responsibility for data processing, providing of course that the DPO is regularly consulted and his advice followed.

Furthermore, it is worth noting that controllership is an element of fact. Any contractual or other arrangements made by the interested parties (e.g. mere designation of X as the controller in a consent form or a contract assigning controllership to B, whereas in fact, A determines the means and purposes of processing) is not binding on the data subject or data protection authorities.[10]

3 Legal grounds for sharing language data and freedom of contract to determine duties

Sharing data constitutes processing of personal data which requires the identification of a legal basis. The suitable legal grounds could be the data subject's consent, public interest research or legitimate interest (GDPR Art. 6 (1) (a), (e), (f)). Legal grounds for conducting research have been previously studied (see, Lindén et al. 2019; Kelli et al. 2019).

A key issue here is how much freedom parties have in determining who has which obligations under the GDPR. According to WP29 "*Being a controller is primarily the consequence of the factual circumstance that an entity has chosen to process personal data for its own purposes*" (2010: 8). WP29 clarifies further that the control could originate from the factual influence and the assessment of contractual relations is helpful since relevant actors often see themselves as facilitators rather than controllers. The contractual terms, however, are not decisive (2010: 11). WP29 (10/2006) found based on a functional influence test that an entity (SWIFT), despite presenting itself as a processor, it was a controller. This demonstrates that any designation of controller/processor which does not correspond to the facts is void.

The GDPR requires the controller to conclude a contract with the processor with several requirements (Art. 28). For instance, this contract is needed when an individual entity/person deposits language data with a CLARIN member. Special attention should be given to the content of this agreement. The GDPR requires that the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of GDPR and ensure the protection of the rights of the data subject (Art. 28 (1)) and the processor can process the personal data only on documented instructions from the controller (Art.28 (3) a). It means that the parties must not only refer to general obligations under the GDPR but sufficiently describe the nature of responsibilities, taking into account the real risks and activities carried out under the specific research project to ensure that all appropriate safeguards are provided (see, Art. 89 (1)). It is

the CLARIN Centre needs to do a due diligence to make sure that the data has been properly collected. If the due diligence is done correctly, the host of the CLARIN Centre may as well become co-controller of the data set.

[9] The solution is likely to differ when it comes to patents – e.g. in Germany, the professors' privilege in patent law was abolished in 2002, and the rights to a patentable invention developed by an academic now belong to his or her institution. Since in the field of language technologies university patents remain rather exceptional, we believe that in the context of this paper an analogy with copyright is more accurate.

[10] WP29 (2010: 9): "…even though the designation of a party as data controller or processor in a contract may reveal relevant information regarding the legal status of this party, such contractual designation is nonetheless not decisive in determining its actual status, which must be based on concrete circumstances".

essential to address the reimbursement of expenses incurred by the performance of the obligations and the payment of remuneration within the contract.

The agreement between the controller and the processor must be concluded in written form, including in electronic form (Art. 28 (9)).[11] The electronic form has been clarified by the European Parliament (2018): "*However, the rules for entering into contracts or other legal acts, including in electronic form, are not set forth in the GDPR but in other EU and/or national legislation. The e-commerce Directive (Directive 2000/31/EC) provides for the removal of legal obstacles to the use of electronic contracts. It does not harmonise the form electronic contracts can take. In principle, automated contract processes are lawful. It is not necessary to append an electronic signature to contracts for them to have legal effects. E-signatures are one of several means to prove their conclusion and terms*".

The joint controllers are not required to enter into such a contract as the controller and the processor. A transparent arrangement between the joint controllers must be agreed upon to comply with the GDPR (Art. 26 (1)), and the essence of this arrangement should be made available to the data subjects (Art. 26 (2)). WP29 (2010: 24) explains it as follows "*Parties acting jointly have a certain degree of flexibility in distributing and allocating obligations and responsibilities among them, as long as they ensure full compliance. Rules on how to exercise joint responsibilities should be determined in principle by controllers. However, factual circumstances should be considered also in this case, with a view to assessing whether the arrangements reflect the reality of the underlying data processing*". The controller's responsibilities must be clearly defined in accordance with actual data processing, and the arrangement must reflect the respective roles and relationships of the joint controllers *vis-à-vis* the data subjects (Art. 26 (2)). Otherwise, as indicated by WP29 (2010: 24), the processing is considered "*unlawful due to a lack of transparency and violates the principle of fair processing*".

4 Sharing language data within CLARIN framework

When it comes to data sharing, we can distinguish two different situations: 1) an external individual or entity deposits LD with a CLARIN member; 2) CLARIN member itself makes LD available.

The first case involves the conclusion of a deposition agreement between the depositor and the CLARIN member. The depositor determines the access and use conditions which makes the depositor the controller under the GDPR. The CLARIN member acts as the processor since it processes personal data on behalf of the depositor.

In the second scenario, the CLARIN member shares LD on its behalf and is the controller.

The main question in both scenarios is whether the sharing of LD leads to joint controllership. According to the GDPR joint controllers jointly determine the purposes and means of processing.[12] They need to determine their respective duties (Art. 26). Joint controllership could be relevant in the case of data sharing. The European Court of Justice (ECJ) has explained that "*a broad definition of the concept of 'controller', the effective and comprehensive protection of the persons concerned, the existence of joint liability does not necessarily imply equal responsibility of the various operators engaged in the processing of personal data.*

On the contrary, those operators may be involved at different stages of that processing of personal data and to different degrees" (C‑40/17 para 70). This means that processing at different stages can result in joint controllership. The court has also maintained that *"a religious community is a controller, jointly with its members who engage in preaching, of the processing of personal data carried out by the latter in the context of door-to-door preaching organised, coordinated and encouraged by that community, without it being necessary that the community has access to those data, or to establish that that community has given its members written guidelines or instructions in relation to the data processing*" (C‑25/17 para 75). It says that it is possible to be a joint controller even without having access to PD.

[11] According to Art. 28 (1). „*Where processing is to be carried out on behalf of a controller, the controller shall use only processors providing sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject.*"

[12] In practice, only 'purposes' are much more important than 'means' for determining the controller, cf. WP29 opinion: "*while determining the purpose of the processing would in any case trigger the qualification as controller, determining the means would imply control only when the determination concerns the essential elements of the means*" (2010: 14). It can be argued that in the CLARIN context, where the 'essential elements of the means' are generally similar and known to everyone (computational analysis), only purposes matter.

From the CLARIN perspective, the proposed agreement structure for transfer of personal data aims to establish a CLARIN Centre as a Data Processor serving the national CLARIN consortium with each of the consortium members, or an external party, as a controller of its data sets. To this end, Finland proposes a CLARIN Framework Deposition Agreement (FADA) with two appendices:

1) the Data Protection Agreement (DAPA), and
2) the Deposition License Agreements (DELA).

The CLARIN FADA establishes a framework of standard deposition rules for data sets that can be communicated by a CLARIN Centre. Individual data sets are added as attachments to the CLARIN FADA keeping only data set specific information in the DELA, which thereby reduces to a 1-page main document for each data set referring to the general conditions in the FADA and four data set specific appendixes:

1) the data identification, description and citation texts,
2) the deposition license conditions with an end-user license agreement template,
3) a list of third-party copyrights or database rights, and
4) the personal data description and the purpose of use of the data set.

Appendixes 3 or 4 may explicitly be left empty if there are no third-party rights or no personal data in the data set.

In the CLARIN infrastructure, there are three main licensing categories dividing language resources into three groups: 1) Publicly available (PUB); 2) For academic use (ACA) and; 3) For restricted use (RES) (for further discussions, see Oksanen et al. (2010) and Kelli et al. (2018). The CLARIN RES licensing category (for restricted use) is suitable for sharing data sets with personal data.

In the suggestions for how to implement the ethical intent of the GDPR in a research setting, Pormeister (2020) recommends that the original controller stays informed about all further use of a personal data set to inform the data subjects about such further use when necessary. The CLARIN RES license requires that data sets not be communicated to a third party by the end-user because a new legitimate end-user can always obtain a copy directly from CLARIN. As the CLARIN Centre remains a mere processor of personal data to communicate such data to research organisations, the original controller will stay informed about all requests for further use of a data set.

If there is a request for using a data set for a research purpose which is not sufficiently compatible with the original purpose of use, the data subjects need to be informed. From a CLARIN perspective, it is a practical question whether the CLARIN Centre as a processor is commissioned to inform the data subjects or the original controller notifies them, and how one goes about informing them in practice, i.e. will personal communication be possible or is a public announcement sufficient.[13]

5 Conclusion

The determination of the controller is essential since the controller has to guarantee compliance with the GDPR. There could be some misconceptions whether individual researchers or the university is the controller. The authors suggest that the university is the controller. Based on the EU case law, it can be assumed that in addition to an organisation, its members can also be regarded as controllers (see C‑25/17 para 75). There is no established practice regarding universities.

Sharing language data within CLARIN requires specific arrangements depending on the nature of relationships. If a third party deposits data with a CLARIN member, then the party is the controller and the CLARIN member is the processor. If a CLARIN member shares data on its behalf, then it is considered the controller itself.

Several contractual arrangements are needed for sharing language data. However, the main reason for having a CLARIN consortium member become the controller of the PD is that researchers are often very mobile. Sometimes they no longer stay in academia, and sometimes researchers pass away. If someone needs access to the data several years down the road, a CLARIN Centre can still carry on with that role. Even if the researcher no longer is in a position personally to grant access to the data, the data is accessible also with a CLARIN Centre in a co-controller position where both a CLARIN Centre and the researcher separately have control of the data.

[13] It seems to depend on whether the data were collected from the data subject (Art. 13 has only one exception to the obligation to provide information) or not (Art. 14 considers impossibility or disproportionate effort).

References

[C‑311/18] Case C‑311/18. Data Protection Commissioner v Facebook Ireland Limited and Maximillian Schrems (16 July 2020). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1598506855221&uri=CELEX:62018CJ0311 (27.8.2020).

[C‑40/17] Case C‑40/17. Fashion ID GmbH & Co. KG *vs*. Verbraucherzentrale NRW eV, interveners: Facebook Ireland Ltd, Landesbeauftragte für Datenschutz und Informationsfreiheit Nordrhein-Westfalen (29 July 2019). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587057502926&uri=CELEX:62017CJ0040 (16.4.2020).

[C‑25/17] Case C‑25/17. Tietosuojavaltuutettu, intervening parties: Jehovan todistajat (10 July 2018). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587066001018&uri=CELEX:62017CJ0025 (16.4.2020).

[C-22/98] Case C-22/98. Criminal proceedings against Jean Claude Becu, Annie Verweire, Smeg NV and Adia Interim NV (16 September 1999). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587999262601&uri=CELEX:61998CJ0022 (27.4.2020).

[C-66/85] Case C-66/85. Deborah Lawrie-Blum *vs*. Land Baden-Württemberg (3 July 1986). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587999733855&uri=CELEX:61985CJ0066 (27.4.2020).

[C-256/01] Case C-256/01. Debra Allonby v Accrington & Rossendale College, Education Lecturing Services, trading as Protocol Professional and Secretary of State for Education and Employment (13 January 2004). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587999966374&uri=CELEX:62001CJ0256 (27.4.2020).

[CNRS guide 2019] CNRS Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte. GUIDE POUR LA RECHERCHE. Available at https://in-shs.cnrs.fr/sites/institut_inshs/files/pdf/guide-rgpd_2.pdf (26.8.2020).

[DP Handbook 2018] European Union Agency for Fundamental Rights and Council of Europe (2018). Handbook on European data protection law 2018 edition. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-coe-edps-2018-handbook-data-protection_en.pdf (29.3.2020).

[Data Protection Policy] Data Protection Policy of the University of Tartu. Available at https://www.ut.ee/en/data-protection-policy (26.8.2020).

[EC] European Commission. What is a data controller or a data processor? Available at https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controller-processor/what-data-controller-or-data-processor_en (15.4.2020).

[EPRS study 2019] European Parliamentary Research Service. How the General Data Protection Regulation changes the rules for scientific research. July 2019. Available at https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634447/EPRS_STU(2019)634447_EN.pdf (27.8.2020).

[European Parliament 2018] European Parliament (2018). Parliamentary questions. Available at https://www.europarl.europa.eu/doceo/document/E-8-2018-003163-ASW_EN.html (27.4.2020).

[EU Charter of Fundamental Rights] Charter of Fundamental Rights of the European Union. 2012/C 326/02. OJ C 326, 26.10.2012, p. 391-407. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (15.4.2020).

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (29.3.2020).

[Kelli et al. 2019] Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramunas Birštonas, Silvia Calamai, Penny Labrpolou, Maria Gavrilidou, Pavel Straňák (2019). Processing personal data without the consent of the data subject for the development and use of language resources. In: Inguna Skadina, Maria Eskevich (Ed.). Selected papers from the CLARIN Annual Conference 2018. Linköping University Electronic Press, 72-82. Available at http://www.ep.liu.se/ecp/159/008/ecp18159008.pdf (23.4.2020).

[Kelli et al. 2018] Aleksei Kelli, Krister Lindén, Kadri Vider, Penny Labropoulou, Erik Ketzan, Pawel Kamocki, Pavel Straňák (2018). Implementation of an Open Science Policy in the context of management of CLARIN

language resources: a need for changes? In: Maciej Piasecki (Ed.). Selected papers from the CLARIN Annual Conference 2017. Linköping University Electronic Press, 102-111. Available at http://www.ep.liu.se/ecp/147/009/ecp17147009.pdf (23.4.2020).

[Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. In: Koenraad De Smedt (Ed.). Selected Papers from the CLARIN Annual Conference 2015. Linköping University Electronic Press, 13-24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (23.4.2020).

[Lindén et al. 2019] Krister Lindén, Aleksei Kelli, Alexandros Nousias, (2019). To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest. Proceedings of CLARIN Annual Conference 2019: CLARIN Annual Conference, Leipzig, Germany, 30 September – 2 October 2019. Ed. K. Simov and M. Eskevich. CLARIN, 56-60. Available at https://office.clarin.eu/v/CE-2019-1512_CLARIN2019_ConferenceProceedings.pdf (23.4.2020).

[Oksanen et al. 2010] Ville Oksanen, Krister Lindén, Hanna Westerlund (2010). Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN' in Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Available at https://helda.helsinki.fi/handle/10138/29359 (24.4.2020).

[Pormeister 2020] Kärt Pormeister (2020). Transparency in Relation to the Data Subject in Genetic Research - an Analysis on the Example of Estonia. Doctoral dissertation. Irene Kull; Jaak Vilo; Katrin Õunap; Barbara Evans (sup). University of Tartu. Available at https://dspace.ut.ee/handle/10062/66697 (26.8.2020).

[Risak & Dullinger 2018] Martin Risak, Thomas Dullinger. The Concept of 'Worker' in EU Law: Status Quo and Potential for Change. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3190912 (27.4.2020).

[WP29 2010] Article 29 Working Party. Opinion 1/2010 on the concepts of "controller" and "processor". Adopted on 16 February 2010. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en.pdf (29.3.2020).

[WP29 2006] Article 29 Working Party. Opinion 10/2006 on the processing of personal data by the Society for Worldwide Interbank Financial Telecommunication (SWIFT). Adopted on 22 November 2006. Available at https://www.dataprotection.ro/servlet/ViewDocument?id=234 (27.4.2020).