CLARIN Annual Conference Proceedings

2022

Edited by

Tomaž Erjavec, Maria Eskevich

10 – 12 October 2022
Prague, Czechia

Please cite as:
CLARIN Annual Conference Proceedings, 2022. ISSN 2773-2177 (online).
Eds. T. Erjavec and M. Eskevich.
Prague, Czechia, 2022.

Programme Committee

Chair:

- Tomaž Erjavec, Jožef Stefan Institute (SI)

Members:

- Starkaður Barkarson, Árni Magnússon Institute for Icelandic Studies (IS)
- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Eva Hajičová, Charles University Prague (CZ)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Neeme Kahusk, University of Tartu (EE)
- Krister Lindén, University of Helsinki (FI)
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- 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)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadič , University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Vincent Vandeghinste, Instituut voor de Nederlandse Taal (Dutch Language Institute), the Netherlands & KU Leuven (BE)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Andreas Witt, University of Mannheim (DE)
- Friedel Wolff, South African Centre for Digital Language Resources, North-West University (ZA)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- 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:

- Olivier Baude, FR
- Federico Boschetti, IT
- Angelo Mario Del Grosso, IT
- Dimitrios Galanis, GR
- Maria Gavriilidou, GR
- Zijian Győző Yang, HU
- Kinga Jelencsik-Mátyus, HU
- Bence Nyéki, HU
- Christophe Parisse, FR
- Valeria Quochi, IT
- Efstathia Soroli, FR
- Thorsten Trippel, DE

CLARIN 2022 submissions, review process and acceptance

- Call for abstracts: 16 December 2021, 21 February 2022
- Submission deadline: 29 April 2022
- In total 21 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 07 and 10 June 2022
- Notifications to authors: 30 June 2022
- 16 accepted submissions

More details on the paper selection procedure and the conference can be found at https://www.clarin.eu/event/2022/clarin-annual-conference-2022.

Table of Contents

Language resources and CLARIN centres

Tools and workflows. Part 1

Tools and workflows. Part 2

Legal Questions

Curation of Language Resources

Research cases

ACTER 1.5: Annotated Corpora for Term Extraction Research

Ayla Rigouts Terryn, Veronique Hoste, Els Lefever
LT3, Language and Translation Technology Team
Department of Translation, Interpreting and Communication – Ghent University
firstname.lastname@ugent.be

Abstract

This contribution presents version 1.5 of the Annotated Corpora for Term Extraction Research (ACTER) dataset. It includes domain-specific corpora in three languages (English, French, and Dutch) and four domains (corruption, dressage (equitation), heart failure, and wind energy). Manual annotations are available of terms and Named Entities for each corpus, with almost 20k unique annotations in total. Significant improvements have been made, most notably the inclusion of sequential annotations. Additionally, an online demo – D-Termine – has been launched for monolingual and bilingual automatic term extraction from parallel corpora, based on the dataset.

1 Introduction and Related Research

Terminology is the specialised vocabulary that defines domain-specific concepts (International Organization for Standardization, 2019). Terms consist of one or more words, which often capture essential information in a concise way. A 2004 Canadian study on *The Economic Value of Terminology* (Champagne, 2004) estimates that "[t]erminology research is required for 4% to 6% of all words in a text" and that translators spend about 20% to 60% of their work on "terminology activities", depending on their experience. To research terminology, translators look to authoritative, comprehensive resources (Kageura and Abekawa, 2013). However, terminological resources are often impossible to keep up to date, do not contain enough information on variation and contexts, and cover only limited domains and language pairs. Therefore, translators resort to unstructured resources: collections of specialised texts and web searches (León-Araúz et al., 2020). Monolingual automatic term extraction (ATE) has been proposed to facilitate this, by automating term identification in specialised texts. The results of ATE, usually a list of unique candidate terms, can be used by language professionals directly, and as preprocessing information for other natural language processing (NLP) tasks, e.g., sentiment analysis (De Clercq et al., 2015).

ATE, like many other tasks in the field of NLP, has seen an evolution in recent years from rule-based (Drouin, 2003; Macken et al., 2013), to machine learning methods (Amjadian et al., 2018; Hätty, 2020; Lang et al., 2021). Many of these machine learning approaches are supervised methodologies, meaning that the systems are trained on annotated data, i.e., corpora in which terms have been (manually) identified. This type of data is invaluable for the development of ATE, both as training and as evaluation data, but it is difficult to acquire. Manual term annotation is a time- and effort-consuming process that generally yields low inter-annotator agreement scores due to its subjectivity. Consequently, only a few datasets for ATE are publicly available and widely used, most notable the ACL RD-TEC 2.0 (Qasemizadeh and Schumann, 2016) and GENIA (Kim et al., 2003). Both have been annotated according to different guidelines, are only available in English and cover a single domain, computational linguistics and biomedicine respectively. While these are valuable datasets, comparisons of ATE performance across languages and domains remains difficult. The ACTER dataset has been designed with these challenges in mind and has been publicly available since 2020 (Rigouts Terryn et al., 2020). The current update to version 1.5 has been released in response to the increasing need for annotations in context, as a consequence of the rise of sequential deep learning methodologies for ATE.

2 ACTER 1.5

ACTER consists of 12 corpora of ±55k tokens in four domains (corruption, dressage (equitation), heart failure, and wind energy), with three languages (English, French, Dutch) per domain. The corpora per domain are comparable (original texts with similar topic and style), except for those in the domain of corruption, which are parallel (aligned translations). Each occurrence of a term or Named Entity has been manually annotated in each corpus based on publicly available guidelines[1]. For terms, a distinction is made between three term labels: Specific Terms (domain-specific and lexicon-specific), Common Terms (domain-specific, not lexicon-specific), and Out-of-Domain Terms (not domain-specific, lexicon-specific). The annotations are available as lists of unique (lowercased) annotations per corpus, including the label. The statistics of the corpora and annotations can be found in Table 1.

CORPUS		**CORPUS COUNTS**			**ANNOTATION COUNTS**				
domain	**lang.**	**docs**	**sentences**	**tokens**	**total**	**Spec**	**Com**	**OOD**	**NE**
corruption	en	12	2,002	50,845	1172	278	641	6	247
	fr	12	1,977	59,130	1207	298	675	5	229
	nl	12	1,988	52,245	1287	308	726	6	247
dressage	en	34	3,090	58,203	1561	769	309	68	415
	fr	78	2,809	61,061	1176	697	234	26	219
	nl	65	3,669	56,450	1541	1020	329	41	151
heart failure	en	190	2,432	55,467	2556	1864	316	157	219
	fr	210	2,177	55,027	2357	1671	486	57	143
	nl	174	2,880	54,966	2215	1535	447	65	168
wind energy	en	5	6,638	57,766	1529	784	295	13	437
	fr	2	4,770	64,989	967	443	308	21	195
	nl	8	3,356	55,328	1229	571	338	21	299
TOTAL		**802**	**37,788**	**681,477**	**18,797**	**10,238**	**5,104**	**486**	**2,969**

Table 1: Counts per corpus of number of documents, sentences, tokens, and number of unique annotations (lowercased) per category, with Spec = Specific Terms, Com = Common Terms, OOD = Out-of-Domain Terms, and NE = Named Entities. These numbers are based on the tokenised corpus and annotations.

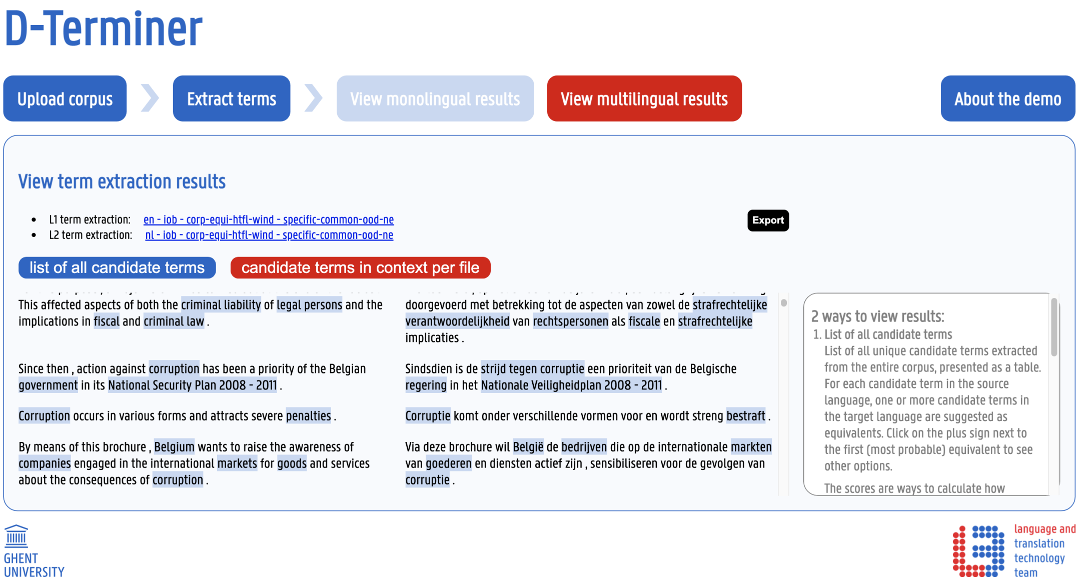
CORPUS		**PROPORTION OF LABELS**		
domain	**language**	**O labels**	**B labels**	**I labels**
corruption	en	81%	12%	7%
	fr	83%	10%	8%
	nl	83%	11%	6%
dressage	en	80%	16%	4%
	fr	83%	13%	4%
	nl	79%	18%	3%
heart failure	en	73%	18%	10%
	fr	79%	13%	8%
	nl	81%	16%	3%
wind energy	en	82%	10%	8%
	fr	83%	9%	7%
	nl	89%	9%	2%
TOTAL		**81%**	**13%**	**6%**

Table 2: Proportions of sequential labels for all tokens per corpus.

[1] http://hdl.handle.net/1854/LU-8503113

For version 1.5, the same annotations (with minor revisions) have been made available in a sequential format, where each token is assigned an IOB (Inside-Outside-Beginning) label. Tokens that are not part of any annotation get an O, the first token of each annotation is labelled B, and each subsequent token that is part of the same annotation is I. Alternatively, the I and B labels can be combined for a binary IO labelling scheme. The conversion of the original annotations to IOB labels has been described in detail in previous work (Rigouts Terryn et al., 2022), which includes a discussion and guidelines on how to fairly evaluate ATE results with the dataset. The latter is important, since ATE is traditionally evaluated using precision, recall, and f1-scores, and calculating these scores based on a list of candidate terms, or based on sequential labels, can lead to very different conclusions. The most notable difference between the original annotations and the IOB ones, is that IOB labelling does not allow nested annotations, so only the longest possible spans of annotations are captured. For instance, the annotation *congestive heart failure* includes the nested annotation of *heart failure*. With IOB labels, only the former was annotated, so the three tokens would be labelled as B, I, I respectively. An overview of the proportion of labels per corpus can be seen in Table 2. Besides the sequential annotations, the original annotations (presented as lists of unique, lowercased terms) are now also available in tokenised form, since the original annotations did not always align with token boundaries. The dataset has been made available as a GitHub repository which can be accessed through the CLARIN portal[2]. It is accompanied by elaborate documentation in the form of a readme.md file and inter-annotator agreement scores can be found in a previous publication (Rigouts Terryn et al., 2020).

3 D-Terminer Online Term Extraction Demo


D-Terminer
Upload corpus
Extract terms
View monolingual results
View multilingual results
About the demo
Export
list of all candidate terms
candidate terms in context per file
GHENT UNIVERSITY
language and translation technology team

View term extraction results

- L1 term extraction: en - iob - corp-equi-htfl-wind - specific-common-ood-ne
- L2 term extraction: nl - iob - corp-equi-htfl-wind - specific-common-ood-ne

This affected aspects of both the criminal liability of legal persons and the implications in fiscal and criminal law .	doorgevoerd met betrekking tot de aspecten van zowel de strafrechtelijke verantwoordelijkheid van rechtspersonen als fiscale en strafrechtelijke implicaties .
Since then , action against corruption has been a priority of the Belgian government in its National Security Plan 2008 - 2011 .	Sindsdien is de strijd tegen corruptie een prioriteit van de Belgische regering in het Nationale Veiligheidplan 2008 - 2011 .
Corruption occurs in various forms and attracts severe penalties .	Corruptie komt onder verschillende vormen voor en wordt streng bestraft .
By means of this brochure , Belgium wants to raise the awareness of companies engaged in the international markets for goods and services about the consequences of corruption .	Via deze brochure wil België de bedrijven die op de internationale markten van goederen en diensten actief zijn , sensibiliseren voor de gevolgen van corruptie .

2 ways to view results:

1. List of all candidate terms
   List of all unique candidate terms extracted from the entire corpus, presented as a table. For each candidate term in the source language, one or more candidate terms in the target language are suggested as equivalents. Click on the plus sign next to the first (most probable) equivalent to see other options.

The scores are ways to calculate how

Figure 1: Screenshot of D-Terminer demo, showing multilingual results in context.

A previously developed supervised machine learning strategy for ATE, using a recurrent neural network with BERT (Devlin et al., 2019) embeddings (Rigouts Terryn et al., 2022), is now available as an online demo[3]. Users can upload a corpus and terminology will automatically be extracted with a model trained on the ACTER dataset. Bilingual ATE from parallel corpora is available as well, so terms can be extracted from a translation memory in each language separately, and then linked cross-lingually using ASTrED word alignment (Vanroy et al., 2021) and frequency ratios. As the methodology relies on multilingual, pretrained BERT embeddings, it can be applied to languages that were not part of the training data.

[2] GitHub: https://github.com/AylaRT/ACTER; CLARIN: http://hdl.handle.net/10032/tm-a2-v4
[3] https://lt3.ugent.be/dterminer/

Therefore, the demo is available in English, French, Dutch, and also German. One of the most notable differences between this demo and similar tools, is that results can be shown in lists, and in the text itself, both for the monolingual and multilingual extractions demonstrated in Figure 1.

4 Conclusion

In this contribution, an updated version of the ACTER dataset was presented, as well as the D-Termine online demo for automatic term extraction. Both are based on previous research and are planned to continue improving based on new research and user feedback.

References

Ehsan Amjadian, Diana Zaiu Inkpen, T. Sima Paribakht, and Farahnaz Faez. 2018. Distributed Specificity for Automatic Terminology Extraction. *Terminology*, 24(1):23–40.

Guy Champagne. 2004. The Economic Value of Terminology: An Exploratory Study. Technical report, submitted to the Translation Bureau of Canada.

Orphée De Clercq, Marjan Van de Kauter, Els Lefever, and Veronique Hoste. 2015. LT3: Applying Hybrid Terminology Extraction to Aspect-Based Sentiment Analysis. In *Proceedings of SemEval 2015*, pages 719–724, Denver, Colorado. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs].*

Patrick Drouin. 2003. Term Extraction Using Non-Technical Corpora as a Point of Leverage. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(1):99–115.

Anna Hätty. 2020. *Automatic Term Extraction for Conventional and Extended Term Definitions Across Domains*. Ph.D. thesis, Universitat Stuttgart, Stuttgart.

International Organization for Standardization. 2019. ISO 1087:2019(en) Terminology work and terminology science — Vocabulary. Technical report.

Kyo Kageura and Takeshi Abekawa. 2013. The Place of Comparable Corpora in Providing Terminological Reference Information to Online Translators: A Strategic Framework. In *Building and Using Comparable Corpora*, pages 285–301. Springer-Verlag, Berlin/Heidelberg, Germany.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA Corpus - a Semantically Annotated Corpus for Biotextmining. *Bioinformatics*, 19(1):180–182.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains. In *Findings of ACL-IJCNLP 2021*, pages 3607–3620, Online. ACL.

Pilar León-Araúz, Arianne Reimerink, and Melania Cabezas-García. 2020. Representing Multiword Term Variation in a Terminological Knowledge Base: A Corpus-Based Study. In *Proceedings of LREC 2020*, pages 2358–2367, Marseille, France. ELRA.

Lieve Macken, Els Lefever, and Véronique Hoste. 2013. TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, 19(1):1–30.

Behrang Qasemizadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In *Proceedings of LREC'16*, pages 1862–1868, Portorož, Slovenia. ELRA.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2020. In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, 54(2):385–418.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2022. Tagging Terms in Text: A Supervised Sequential Labelling Approach to Automatic Term Extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28(1).

Bram Vanroy, Orphée De Clercq, Arda Tezcan, Joke Daems, and Lieve Macken. 2021. Metrics of Syntactic Equivalence to Assess Translation Difficulty. In Michael Carl, editor, *Explorations in Empirical Translation Process Research*, volume 3, pages 259–294. Springer International Publishing, Cham.

Linguistic Autobiographies. Towards the Creation of a Multilingual Resource Family

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

Rosalba Nodari
Università di Siena, Italy
rosalba.nodari@unisi.it

Claudia Soria
CNR-ILC, Italy
claudia.soria@ilc.cnr.it

Alessandro Carlucci
University of Bergen, Norway
alessandro.carlucci@uib.no

Abstract

This paper describes a project aimed at adding a new type of corpus to the CLARIN resource family tree, called 'linguistic autobiographies'. In a linguistic autobiography the writer explicitly reflects on the relationship between him/herself and the language. This genre is fruitfully used in different educational settings, and research has shown that it helps to uncover the social, affective, and psychological dimensions of language learning. The potential of a multilingual collection is discussed starting from Italian data.

1 CLARIN Resource families and linguistic diversity

The CLARIN Resource Family (Fišer et al. 2018) is a user-friendly overview per data type of available language resources in the CLARIN infrastructure aimed at the needs of researchers from digital humanities, social sciences and human language technologies. Resource families are provided according to modality (spoken, multimodal, computer-mediated), genre (historical, academic, literary, newspaper, …), multilingualism, and intended use (reference, L2 learners). These groupings of corpora, lexical resources and tools are meant to facilitate comparative research: for each resource family, a brief description is provided followed by a list of the resources belonging to the family, together with the most important metadata (name, size, annotation, license, language, description and availability). Thus, resource families provide a curated view of the available CLARIN resources. Over the years, this has proven to be a highly visible initiative appreciated by a broad spectrum of CLARIN users (Leonardič and Fišer 2020) and it therefore deserves to be maintained and enlarged. In this paper, we introduce a new textual genre (linguistic autobiographies) and we argue that a resource family devoted to this genre could be useful for both first and second language research and pedagogy, as well as for a better understanding of the linguistic landscape in European schools and universities, as it is shown in the following section.

2 The underutilised potential of linguistic autobiographies in shaping the European multilingual landscape

2.1 What are linguistic autobiographies

A linguistic autobiography is a non-fictional genre where the writer explicitly reflects on the relationship between him/herself and the language. In this self-reflective writing practice, language becomes the overarching organising principle for retracing salient moments in the writer's life. The idea behind this genre is that the acquisition and the interaction of different languages can be seen as the acquisition of selfhood (Ramsdell 2004). Linguistic autobiographies can be considered as both research and a pedagogical tool. They are used by professors and teachers in secondary schools and university classrooms to help students in developing their metalinguistic and metapragmatic abilities; within superdiverse multilingual classrooms, linguistic autobiographies allow students to narrate their multilingual selves and they help make their languages more visible. Linguistic autobiographies can also help linguists in gaining access to language ideologies and attitudes towards language varieties: in particular, these narratives can help understand how ideologies about languages can have an impact on the linguistic behaviour of speakers.

2.2 Linguistic autobiographies as a powerful teaching tool

Linguistic autobiographies are highly versatile in that they can easily be collected without requiring specific skills or academic knowledge. Several templates are available to facilitate the production of linguistic autobiographies, which offer some suggestions to think about languages and self-reflect on key points of the writers' own life (cf. Canobbio 2006; D'Agostino 2007; Luppi, Thüne 2022). For example, one possible template requires mentioning:

1) Family members and personal data (place of birth, eventual relocations, etc.);
2) Family linguistic background: L1 of the grandparents, L1 of the parents;
3) Family linguistic situation: parents and grandparents' linguistic preferences (they speak which language to whom and when); which languages are used for ordinary communication between family members (with children, with the rest of the family); family choices in linguistic education (which language is taught to children); which language is spoken at home; which other varieties are used at home, and for what need (communicative, expressive, affective needs, identity stances, etc.);
4) Family and school attitudes and behaviours: repression of non-standard varieties; are any specific varieties preferred to other varieties? Are there any disfavoured languages? Are there any forbidden languages used in secret between friends or family members?
5) Meeting with linguistic diversity (holiday trips to different regions, community of practices, peer groups, school environment, extended family etc.). Formal and informal language learning (foreign language schools, friends from abroad etc.); are different languages used with different groups of people? Are non-standard varieties used for performing specific identities? Etc.
6) Personal evaluation of language learning in and out of school; ability to perceive different linguistic varieties, social evaluation of accents, stigma and stereotypes towards accents, varieties, languages, etc.

The template (and the terminology used) can be adapted depending on the age, educational background and other characteristics of those involved. In any case, linguistic autobiographies are deeply personal texts that outline the writers' own thoughts and feelings, giving them the possibility for spontaneous expression.

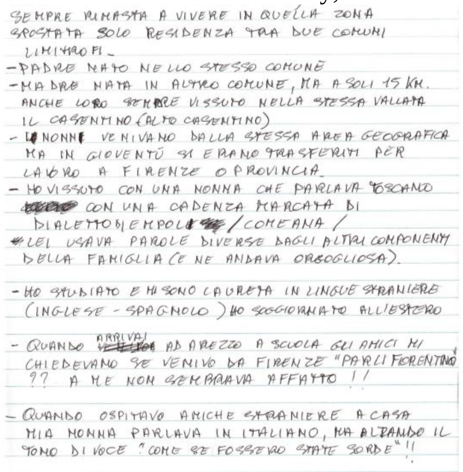
2.3 The societal potential of linguistic autobiographies

Linguistic autobiographies are fruitfully used in different educational settings, and research has shown that this tool helps to uncover the social, affective, and psychological dimensions of language learning (Franceschini, Miecnikowski 2004; Groppaldi 2010; Salvadori, Blondeau, Polimeni 2020). This kind of narrative permits students to develop an awareness of cultural and linguistic diversity, and to learn about the social value of languages. In superdiverse settings, linguistic autobiographies help students in understanding mechanisms of stereotyping and linguistic discrimination and are considered an empowering tool. Teachers can also gain access to the students' linguistic learning process, and they can discover students' learning practices, and reasons for studying languages, as well as their needs and expectations. For policy makers and stakeholders, linguistic autobiographies can help in understanding students' motivation for language learning, thus developing specific school curricula for addressing students' communicative needs. For those who are interested in the sociology of languages, linguistic autobiographies can also provide unique information about informal language-learning opportunities and the different values that speakers attach to different types of multilingualism.

3 Contributing linguistic autobiographies to CLARIN

The VLO repository already offers a selection of linguistic autobiographies collected in two (non-exclusively) Italian-speaking settings, namely Language Biographies from South Tyrol and from Basel. However, these two corpora consist of audio interviews, together with their transcriptions. The aim of this proposal is, instead, to create a new multilingual corpus that comprises several written linguistic autobiographies of both L1 and L2 speakers, collected in different languages and different national settings. This new corpus will offer comparable corpora of the same genre with the appropriate meta-data profile. For example, by searching the corpus according to speakers' L1, it will be possible to compare the biographies of speakers with the same L1 across different countries, educational settings, etc.

Firstly, a selection of linguistic autobiographies collected in Italy and Norway in different educational settings will be deposited in the CLARIN repository. Currently, the Italian corpus comprises almost 200 linguistic autobiographies collected during university linguistic courses, ca. 50 autobiographies of secondary school students and ca. 40 autobiographies of secondary school teachers (see a specimen in Fig. 1). The data will be digitised, anonymised, and the appropriate metadata description will be chosen. The metadata profile used for the Italian corpus will then be applied to the different collections. The anonymisation and metadatation will give the possibility to make the corpus downloadable and accessible under public licences. Together with linguistic autobiographies, a multilingual template will be made available to the CLARIN community, in order to facilitate the collection of linguistic autobiographies.



SEMPRE RIMASTA A VIVERE IN QUELLA ZONA
SPOSTATA SOLO RESIDENZA TRA DUE COMUNI
LIMITROFI.
-PADRE NATO NELLO STESSO COMUNE
-MADRE NATA IN ALTRO COMUNE, MA A SOLI 15 KM.
ANCHE LORO SEMPRE VISSUTO NELLA STESSA VALLATA
IL CASENTINO (ALTO CASENTINO)
- I NONNI VENIVANO DALLA STESSA AREA GEOGRAFICA
MA IN GIOVENTÙ SI ERANO TRASFERITI PER
LAVORO A FIRENZE O PROVINCIA.
- HO VISSUTO CON UNA NONNA CHE PARLAVA TOSCANO
CON UNA CADENZA MARCATA DI
DIALETTO DI EMPOLI /COMEANA /
*LEI USAVA PAROLE DIVERSE DAGLI ALTRI COMPONENTI
DELLA FAMIGLIA (E NE ANDAVA ORGOGLIOSA).

- HO STUDIATO E MI SONO LAUREATA IN LINGUE STRANIERE
(INGLESE - SPAGNOLO) HO SOGGIORNATO ALL'ESTERO

- QUANDO ARRIVAI AD AREZZO A SCUOLA GLI AMICI MI
CHIEDEVANO SE VENIVO DA FIRENZE "PARLI FIORENTINO"
?? A ME NON SEMBRAVA AFFATTO !!

- QUANDO OSPITAVO AMICHE STRANIERE A CASA
MIA NONNA PARLAVA IN ITALIANO, MA ALZANDO IL
TONO DI VOCE "COME SE FOSSERO STATE SORDE"!!

Figure 1. An example of linguistic autobiography from the University of Siena corpus

The creation of a new linguistic corpus in the CLARIN resource family initiative does carry the need for further consideration regarding metadata description. According to Leonardič and Fišer (2020), this is a crucial issue in building resource families. The curated resources tend to have different depths and breadths of metadata description, which in turn has consequences on their final usability. It is of utmost importance, therefore, that linguistic autobiographies are described in such a way that their collection under a resource family is straightforward and allows for their maximum comparability. We believe that a metadata description that is adequate for building resource families must satisfy two interacting main requirements: a) exhaustiveness and b) comparability.

Exhaustive metadata description is obviously important not only for the sake of accuracy, but also in order to maximise the impact on the traceability of linguistic autobiographies and the possibility for them to be discovered and included in future collections. In order to ensure the highest possible degree of comparability with other resources, either already present or to be added in the future, metadata description should also take into account the metadata sets used for describing resources that are partially overlapping for content and/or genre with linguistic autobiographies. It is likely that linguistic autobiographies have features in common with already existing or future resources (such as resources belonging to oral histories).

From a first analysis of the VLO we have identified oral interviews, general autobiographies and personal narratives as the most similar genres already represented in CLARIN collections. Linguistic autobiographies, on the other hand, show some peculiar features such as a) the written modality vs.

mainly oral one and b) their strong emphasis on the linguistic component: language is the key around which the narrative is built and articulated, and the recollection of one's life follows a linguistic path, while in general oral narratives focus on the main events in the lives of speakers.

4 Towards a CLARIN Resource Family for Linguistic Autobiographies

We are confident that in some of the countries involved in the CLARIN network, linguistic autobiographies are already used in school and university settings. With this project, thus, we would like to help uncover this eventually already existing material as well as to encourage the production of new one. A multilingual collection of such written material will indeed offer an invaluable picture from several perspectives. Firstly, it can be used as teaching material, from school classes of any grade to university courses, in order to raise awareness of heritage languages, accentism and glottophobia. Secondly, it can help teachers to better understand the most used and known languages in the classrooms. In addition, it represents a useful tool to verify among pupils, students, and teachers, the pervasiveness of the concept of linguistic error and deviation in describing linguistic repertoires.

Comparable corpora of linguistic autobiographies will also provide valuable quantitative and qualitative data to researchers interested in a variety of topics – such as language attitudes, language and migration, multilingualism and language contact. Finally, this new resource can help policymakers in designing linguistic policies more consistent with the different linguistic landscapes which are truly present in different European schools and universities.

References

Canobbio, S. 2006. Dialetto dei giovani e politiche linguistiche delle famiglie, appunti dal Piemonte. In: Marcato, G.: *Giovani, lingue e dialetti, Atti del Convegno,* Sappada - Plodn, 29 June - 3 July 2005. Unipress, Padova: 239-244.

D'Agostino, M. 2007. *Sociolinguistica dell'Italia contemporanea.* Il Mulino, Bologna.

Fišer, D., Lenardič, J., and T. Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 1320–1325.

Franceschini, R., and Miecnikowski, J. (eds) 2004. *Leben mit mehreren Sprachen. Vivre avec plusieurs langues. Sprachbiographien. Biographies langagières*. Lang, Bern.

Groppaldi, A. 2010. L'autobiografia linguistica: strumento per una moderna didattica dell'italiano L2-LS. *Italiano LinguaDue,* 2(1): 89-103.

Lenardič, J. and D. Fišer. 2020. Extending the CLARIN Resource and Tool Families. In: Navarretta, C. & Eskevich, M.: *Proceedings of CLARIN Annual Conference 2020,* 05-07 October 2020, Virtual Edition: 1–5.

Luppi, R, and Thüne, E.-M. (eds) 2022. Biografie linguistiche. Esempi di linguistica applicate, Centro di Studi Linguistico-Culturali (CeSLiC), Bologna.

Ramsdell, L. 2004. Language and Identity Politics: The Linguistic Autobiographies of Latinos in the United States. *Journal of Modern Literature* 28(1): 166-176.

Salvadori, E.; Blondeau, N.; Polimeni, G. (eds.) 2020. La formazione e le competenze di un insegnante riflessivo. *Italiano LinguaDue,* 12(2): 352-389.

CLARIN-LV: Many Steps till Operation

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

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

Roberts Darģis
Institute of Mathematics and Computer Science
University of Latvia
roberts.dargis@lumii.lv

Eduards Lasmanis
Institute of Mathematics and Computer Science
University of Latvia
eduards.lasmanis@lu.lv

Arnis Voitkāns
Institute of Mathematics and Computer Science
University of Latvia

arnis.voitkans@lu.lv

Abstract

Inspired by the previous submissions from CLARIN national consortia, in this abstract we summarize the most important steps and achievements during the implementation phase of CLARIN-LV research infrastructure. Although CLARIN-LV was an active supporter of CLARIN goals during the preparatory phase, Latvia joined CLARIN ERIC only in 2016. During the last five years CLARIN-LV became an active C-center, supporting and collaborating with digital humanities, developing and sharing language resources developed by Latvian academic community, as well as active contributor and participant of CLARIN international activities.

1 Introduction

Latvia has been participating in the CLARIN initiative since its beginning (Skadiņa, 2009). Although Latvia was among countries that signed Memorandum of Understanding, Latvia joined CLARIN ERIC only in 2016.

Since 2018, activities of CLARIN-LV are supported through the targeted project “University of Latvia and institutes in the European Research Area – Excellency, activity, mobility, capacity” (Skadiņa et al., 2020). The objectives of the project include consortium building and repository set up. Additional funding is provided through the State Research programme “Digital resources for humanities: integration and development” and the State Research Programme "Letonika – Fostering a Latvian and European Society" project "Research on Modern Latvian Language and Development of Language Technology”. The moderate funding we receive from the targeted project and some satellite projects allows us to concentrate our activities around three main dimensions:

- Repository building and filling;
- Consortium building and collaboration with digital humanities community;
- Education and teaching.

Besides national activities, CLARIN-LV supports regional cooperation (e.g. bi-annual Baltic HLT conferences and Baltic Summer Schools in Digital Humanities) and participates in CLARIN ERIC activities (e.g., CLARIN Resource Families, ParlaMint (Erjavec et al., 2022) and CLARIN for Teachers).

2 Repository

Creation, maintenance and sharing of language resources has been among priorities of the Institute of Mathematics and Computer Science (IMCS) of the University of Latvia since 1988, when the Latvian folk songs and the first Latvian Bible were digitized. Therefore, after joining CLARIN ERIC, one of our first tasks was to set up a CLARIN-LV repository[1]. After analysis of different options used by other CLARIN centres, LINDAT/CLARIN repository system[2] was chosen. The main reasons for this choice were availability of installation resources and documentation, documented and automated maintenance, approval by many CLARIN centres, reliability, and technical support.

CLARIN-LV repository was set up in March, 2020, shortly before COVID-19 crisis started. Thus we were very lucky to introduce our repository to the humanities researchers at the first national CLARIN conference[3]. During the conference national coordinators of FIN-CLARIN and Centre of Estonian Language Resources introduced to their CLARIN repositories, highlighted interesting language resources and presented user involvement activities.

CLARIN-LV repository mostly focuses on Latvian (and Latgalian) language resources, but not excluding other languages, especially morphologically rich languages. In the beginning the repository catalogued about 10 language resources, mostly developed IMCS. Today CLARIN-LV provides information about 36 language resources and tools (23 corpora, 11 lexical conceptual resources and 2 tools), developed not only by IMCS, but also by partner institutions: the Institute of Latvian Language (University of Latvia), the Institute of Literature, Folklore and Art (University of Latvia), Rēzekne Academy of Technologies, Ventspils University of Applied Sciences and Latvian Language Agency. Figure 1 summarizes information about most viewed items in CLARIN-LV repository.

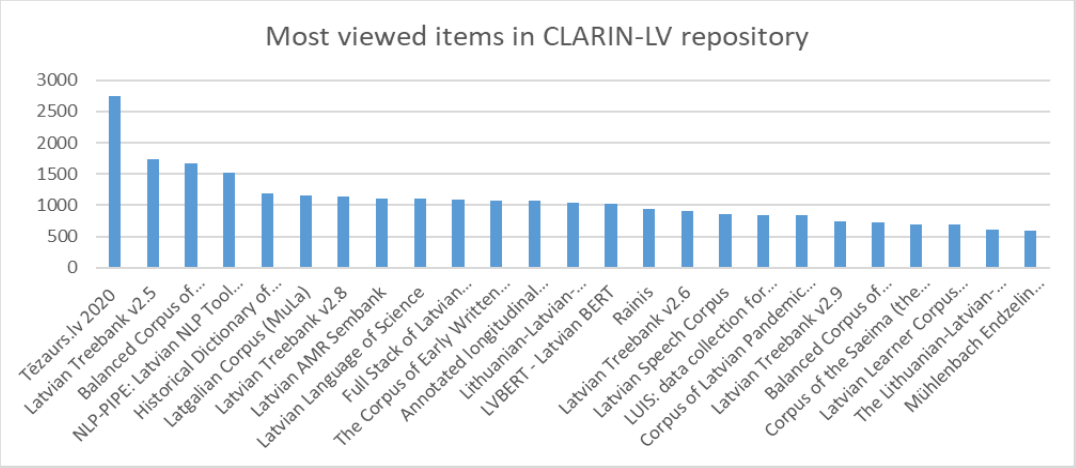


Figure 1: Top 25 most viewed language resources and tools in CLARIN-LV repository.

2.1 Corpora

CLARIN-LV repository lists 23 corpora – 21 text and 2 audio (Saulīte et al, 2022). Most of the corpora are monolingual Latvian (19), one corpus represents Latgalian language. These corpora are available mostly for browsing and querying through NoSketch Engine (Kilgarriff et al., 2014; Rychly, 2007), however, they are not available for download due to the copyright restrictions.

Modern Latvian is primarily represented through the Balanced Corpus of Modern Latvian LVK2018 (Levāne-Petrova and Darģis, 2018). Other widely used corpora are Latvian Treebank, Latgalian corpus and FullStack.

While Latvian text data are rather well represented, access and availability of speech corpora is limited. Currently CLARIN-LV lists only two corpora – Latvian Speech corpus LaRKo and annotated longitudinal Latvian children's speech corpus LAMBA.

2.2 Lexical Resources

CLARIN-LV repository lists 11 lexical conceptual resources, most of them are monolingual Latvian. *Tezaurs.lv* (Spektors et al, 2019) is the largest open lexical dataset and on-line dictionary for Latvian.

[1] https://repository.clarin.lv/repository/
[2] GitHub - clarinsi/clarin-dspace: LINDAT/CLARIN digital repository based on DSpace
[3] Konference "Latviešu valodas digitālie resursi un rīki vienotā pētniecības infrastruktūrā" (clarin.lv)

The dictionary is popular not only among researchers, but also widely used by the general public – translators, journalists, students and many others, receiving more than 80,000 requests per day. It is the most viewed item of CLARIN-LV repository, besides browsing and querying, it is also available for download. Other popular lexical conceptual resources are Historical Dictionary of Latvian, database of Latvian Language of Science and Lithuanian-Latvian-Latgalian dictionary.

2.3 Tools

Tools are not so well represented in the CLARIN-LV repository. The most popular (4th among CLARIN-LV language resources) is Latvian NLP tool pipeline NLP-PIPE (Znotiņš, 2015), which integrates various open-source Latvian NLP tools. A pipeline supports text tokenization and sentence splitting, morphological tagging, named entity recognition, syntactic parsing, semantic parsing, etc. Latvian BERT is also available for download from the repository.

2.4 Collaboration: Case of ParlaMint

CLARIN gives great opportunities for collaboration. One of such cooperation projects is ParlaMint corpora of parliamentary proceedings. In this project researchers from multiple countries encoded their parliamentary corpus data in a unified format which makes them more accessible for other researchers and easily integrable in different platforms, such as SketchEngine[4]. CLARIN-LV contributed to this initiative with Latvian parliamentary corpus (Darģis et al., 2018), created before this initiative.

3 Support for Digital Humanities

Serving Digital Humanities is one of the main goals of CLARIN research infrastructure. CLARIN-LV implements this mission through several activities – targeted practical workshops, collaboration through State Research programme projects, and supporting and distributing knowledge about Latvian language resources and tools (LRTs) through Knowledge Center for Systems and Frameworks for Morphologically Rich Languages SAFMORIL (Axelson et al., 2021).

Many SSH researchers in Latvia still have insufficient knowledge on how NLP techniques can help them to solve specific research questions. CLARIN-LV together with Digital Humanities initiative[5] regularly organizes conferences and seminars[6] to introduce various language resources and tools and to demonstrate their use for research. While conferences provide a good overview of research and development activities, practical seminars are more targeted towards problem solving. They are organized in small groups, usually cover one specific topic and aim at active involvement of each participant.

CLARIN-LV together with digital humanities researchers is involved in several research and development projects financed by the Latvian Council of Science. These projects aim to adopt LRTs for digital humanities use cases. For example, NLP-PIPE is currently being tested in two selected use cases: the named entity recognition service is being integrated into the Latvian literature platform *Literatura.lv* allowing automatic recognition of person names, place names, and events mentioned in texts, while NLP-PIPE tools for morphological annotation are being applied on the text analysis platform of the Latvian National Digital Library[7].

4 Consortium

When Latvia joined CLARIN ERIC, the Ministry of Education and Science appointed IMCS to represent Latvia in CLARIN ERIC and to lead the CLARIN consortium in Latvia.

As the leading language technology research organization in Latvia, IMCS has established long term cooperation with many research organizations in Latvia, involved in development or usage of language resources and tools for research and education. This collaboration led to a formal consortium agreement. The CLARIN-LV consortium partners are universities and research institutions: University of Latvia,

[4] ParlaMint in SketchEngine: https://www.sketchengine.eu/parlamint-corpora-of-parliamentary-debates/

[5] digitalhumanities.lv

[6] Presentation form conferences and seminars are published at *clarin.lv* website: Konferences un semināri (clarin.lv)

[7] https://lndb.lv

Institute of Literature, Folklore and Art of the University of Latvia, Rīga Stradiņš University, Liepaja University, Rezekne Academy of Technologies, National Library of Latvia and IMCS UL.

5 Teaching

Since Latvia is a rather small country, teaching of Computational Linguistics and Digital Humanities courses almost automatically involves introduction to CLARIN-LV catalogue of language resources and tools. Today most of teaching activities concentrate in Faculty of Humanities of the University of Latvia[8]. Different courses for master and doctoral students involve use of corpora from CLARIN-LV and *korpuss.lv* websites. In targeted seminar, students are asked to explore CLARIN VLO, select tool or resource and present it at seminar.

In autumn, 2022, a new course "Language Corpora in the Educational Process" in the professional program for teachers is offered by the Faculty of Pedagogy, Psychology and Art of the University of Latvia. This course aims to develop the future teachers' ability to use corpora for research of social and regional variants of the Latvian language, acquisition of Latgalian written language, research of vocabulary and expansion of vocabulary, as well as teaching topics of sociolinguistics and pragmatics in Latvian language lessons.

6 Conclusion

In this abstract we summarized CLARIN-LV activities and highlighted most important steps to make the national node operational. Being among the smallest countries of CLARIN infrastructure and joining CLARIN only in 2016, has some positive and some negative consequences. On one hand, many administrative and technical issues were solved when we joined, thus we can use reliable and stable solutions developed by the consortium. On the other hand, being among the newest members of CLARIN, clearly revealed a gap in needs, understanding and readiness to serve the SSH community.

While the targeted project for setting up CLARIN-LV node will end in December, 2022, CLARIN-LV plans to continue activities through several language technology related projects. Recently started project "Research on Modern Latvian Language and Development of Language Technology" of the State Research Programme “Letonika – Fostering a Latvian and European Society" will continue supporting cooperation with digital humanities researchers, the development of new resources and their subsequent inclusion in the CLARIN-LV repository. Additional funding is also expected from the Latvia's Recovery and sustainability plan[9], which envisions activities for the development and implementation of high-level skills in language technology.

Acknowledgements

Activities of CLARIN-LV are supported by project of the European Regional Development Fund Nr. 1.1.1.5/18/I/016 University of Latvia and institutes in the European Research Area - Excellency, activity, mobility, capacity and in a synergy with the National Research Programmes *Digital Resources of the Humanities* (grant agreement No. VPP-IZM-DH-2020/1-0001), and *Letonika – Fostering a Latvian and European Society* (grant agreement No. VPP-LETONIKA-2021/1-0006).

References

Axelson, E., Moshagen, S., Vaičenonienė, J., Skadina, I., Lindström Tiedemann, T. and Lindén, K. 2021. Tour de CLARIN: CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages (SAFMORIL).

Dargis, R., Auzina, I., Bojars, U., Paikens, P., Znotins, A. 2018. Annotation of the Corpus of the Saeima with Multilingual Standards. Proceedings of the 2018 ParlaCLARIN Workshop, 2018.

Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. 2022. *The ParlaMint corpora of parliamentary proceedings*. Lang Resources & Evaluation.

[8] Riga Technical University also offers Master program in Digital Humanities

[9] https://likumi.lv/ta/id/322858-par-latvijas-atveselosanas-un-noturibas-mehanisma-planu

Levāne-Petrova, K. and Darģis, R. 2018. *Balanced Corpus of Modern Latvian (LVK2018)*, CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/11.

Kilgarriff, A., Baisa, V., Bušta, J. et al. 2014. *The Sketch Engine: ten years on*. Lexicography, 1(1), 7-36.

Rychly, P. 2007. Manatee/Bonito-A Modular Corpus Manager. RASLAN, 65-70.

Saulīte, B., Dargis, R., Gruzitis, N., Auzina, I., Levāne-Petrova, K., Pretkalniņa, L. 2022. *Latvian National Corpora Collection – Korpuss.lv*. LREC 2022 proceedings (in press).

Skadiņa, I., Auziņa, I., Grūzītis, N. and Znotiņš, A. 2020. *Clarin in Latvia: From the preparatory phase to the construction phase and operation.* Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries (DHN).

Skadiņa I. 2009. *CLARIN in Latvia: current situation and future perspectives*. Proceedings of the NODALIDA 2009 workshop Nordic Perspectives on the CLARIN Infrastructure of Common Language Resources, May 14, 2009, Odense, Denmark, NEALT Proceedings Series, Vol. 5 (2009), 33-37.

Spektors, A. et al. 2019. *Tēzaurs.lv 2020*, CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/9 .

Znotiņš, A. 2015, *NLP-PIPE: Latvian NLP Tool Pipeline*, CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/4.

BabyLemmatizer: A Lemmatizer and POS-tagger for Akkadian

Aleksi Sahala
University of Helsinki, Finland
aleksi.sahala@helsinki.fi

Tero Alstola
University of Helsinki, Finland
tero.alstola@helsinki.fi

Jonathan Valk
University of Helsinki, Finland
jonathan.valk@helsinki.fi

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

Abstract

We present a hybrid lemmatizer and POS-tagger for Akkadian, the language of the ancient Assyrians and Babylonians, documented from 2350 BCE to 100 CE. In our approach the text is first POS-tagged and lemmatized with TurkuNLP trained with human-verified labels, and then post-corrected with dictionary-based methods to improve the lemmatization quality. The post-correction also assigns labels with confidence scores to flag the most suspicious lemmatizations for manual validation. We demonstrate that the presented tool achieves a Lemma+POS labeling accuracy of 94%, and a lemmatization accuracy of 95% in a held-out test set.

1 Introduction

Application of computational methods to historical text corpora provides interesting opportunities for studying large-scale phenomena that are difficult to perceive through close reading of texts. This often requires careful normalization of the language, because in many past societies spelling conventions were not fully standardized, and the corpora can contain documents written in several synchronic and diachronic variants of the language. The languages can also be morphologically complex, which further complicates even such fundamental tasks as searching for all attestations of a certain word in the corpus.

One way to normalize historical languages is lemmatization, which labels words with their dictionary forms regardless of their morphology and spelling. In this paper, we present a lemmatizer for Akkadian, an extinct language that was widely used in ancient Mesopotamia.

The motivation for this tool emerges from close co-operation between the FIN-CLARIN coordinated Language Bank of Finland and the Centre of Excellence in Near Eastern Empires, a University of Helsinki-based research project focusing on the study of the Near East in the first millennium BCE. As part of this co-operation, the Language Bank of Finland collects corpora of ancient Mesopotamian texts written in the Akkadian language in the Korp concordance service.[1] Korp offers several useful functionalities for historians from flexible search options to generating statistics from text metadata.(Borin et al., 2012)

At present, Korp hosts a version of the Open Richly Annotated Cuneiform Corpus (Oracc),[2] which comes with human-verified lemmatization. The next Akkadian corpus to be included in Korp is Achemenet,[3] which has not been lemmatized. The only Akkadian lemmatizer currently available (Tinney, 2019) requires extensive human supervision. To minimize the need for human intervention, our aim is to lemmatize the Achemenet corpus by first training the TurkuNLP's lemmatizer using the available Oracc data, and then applying simple dictionary-based post-correction scripts.

2 The Akkadian Language

Akkadian is an extinct East Semitic language documented in hundreds of thousands of cuneiform tablets excavated across the Near East. The earliest written exemplars of Akkadian date back to the Sargonic Pe-

[1] http://korp.csc.fi

[2] http://oracc.org/

[3] http://www.achemenet.com/

riod (2350-2170 BCE), after which the language is mostly documented in its two main dialects: Assyrian (1950-600 BCE) and Babylonian (2100 BCE - 100 CE) (von Soden, 1995).

Like other Semitic languages, Akkadian morphology employs nonconcatenative root-pattern morphotactics in stem formation and concatenative morphotactics in the attachment of various grammatical affixes to the stems. For example, the verbal form *ludlul* ”let me praise (it)!” consists of the first person singular precative suffix {lu} attached to the preterite stem {dlul}, which is formed from the root *dll* of the verb *dalālu* ”to praise”. Although the morpheme boundaries are transparent in this example, various morphophonological processes often obscure the underlying structure of the word, complicating recognition of the root radicals (von Soden, 1995).

Another layer of complexity emerges from the cuneiform script that developed to represent the linguistically unrelated Sumerian language before being adopted to represent Akkadian in the 24th century BCE. Although the Akkadian language was generally written syllabically, scribes sometimes favoured the use of Sumerian logograms, especially in certain genres of text. The Akkadian verbal form *iddin* ”(s)he gave it”, can for instance, be spelled syllabically as *id-din, i-din, id-di-in* or *i-di-in*, but logographic or logo-syllabic spellings like SUM and SUM-*in* are also attested.

In Akkadian transliteration, logograms are represented in capital letters and named after their base reading values in Sumerian rather than Akkadian. For this reason, the character level relationship between the graphemic and phonemic forms of logographic spellings is typically suppletive. Many logograms are also ambiguous and can have different readings in different contexts. For example, the Sumerian logogram IGI (depicting an eye) can indicate any form of the words *īnu* ”eye”, *pānu* ”front”, *mahru* ”before” and *amāru* ”to see”.

2.1 Digital Resources

For an extinct language, Akkadian is fairly well resourced, and texts comprising about 3-4 million tokens in total have been digitized.[4] Some larger text corpora are Oracc (the Open Richly Annotated Cuneiform Corpus) with 19,000 Akkadian texts, Achemenet with 3,000 texts and Archibab with 10,000 texts. A complete survey of Akkadian digital resources is given in Charpin (2014).

3 Previous Work

Due to the previously discussed complexity of the Akkadian morphology and script, lemmatization is considered a mandatory step into making any digital corpus of Akkadian searchable or suitable for computational analysis (Maiocchi, 2019). To date, however, only Oracc provides extensive lemmatization for Akkadian texts, totalling about 1.5 million lemmatized words. Oracc is lemmatized using a dictionary-based tool known as L2 (Tinney, 2019), which populates new texts with lemmata and POS-tags based on a labelled glossary extracted from previously lemmatized texts. Texts are then checked manually word-by-word, filling in lemmata for out-of-vocabulary words and resolving possible ambiguities.

4 Description of BabyLemmatizer

BabyLemmatizer combines the use of neural networks and dictionary-based lemmatization. The backbone of our tool is Turku Neural Parser Pipeline (TurkuNLP) (Kanerva et al., 2018), a state-of-the-art neural lemmatizer and POS-tagger, for which we train a model using Oracc data. In the lemmatization process, we first provide the text with POS-tagging and raw lemmatization with TurkuNLP and then apply post-corrections to the result to improve lemmatization accuracy. Our post-correction involves three steps:

The first step overrides all predictions for in-vocabulary words. This minimizes the effect of mislearned character level relationships between spellings and their lemmata. We calculate the degree of ambiguity for all lemmatizations in the training data and create a *master glossary* of word forms that have a low degree of ambiguity. We then override all lemmatizations of in-vocabulary words using this data. The degree of ambiguity for a word form is considered to be low, if any lemma+POS label constitutes over N percent of all the labels assigned to it in the training data. Based on our experiments, an N-value

[4] This is our crude estimate based on the number of texts listed in various text corpora.

of 60% seems to produce consistently good results. We leave highly ambiguous lemmata as they are. The second step aims to assign correct lemmata to ambiguous word forms, especially logograms. Here we calculate co-occurrence probabilities for lemmata and their adjacent POS-tags in the training data, and then assign the most likely lemmata for all word forms in the text. We rely on POS-tags instead of surrounding lemmata due to the very reliable POS-tagging accuracy of TurkuNLP. This step allows us to reconfirm that our close-to-unambiguous lemmata are likely correct, and that the ambiguous word forms are lemmatized with the most likely option. The minimum probability threshold is adjustable, but in our experiments we always accept the most likely lemma in the given context.

Finally, we apply various other post-corrections to the data, such as removing the lemmatization from numbers and words that occur in badly damaged sections of the tablet (unreadable signs are indicated in transliteration with x, as in *x-x-in-nu*, which makes them easy to find). This is done to make the lemmatizations more consistent with Oracc conventions and to prevent TurkuNLP from attempting to predict reconstructions that are beyond human comprehension. We also heuristically detect some obvious lemmatization errors, such as verbs that show impossible or very unlikely dictionary form patterns. Nonetheless, these can only be flagged, but not fixed automatically.

4.1 Confidence scoring

Post-processing also assigns lemmatizations with confidence scoring that helps Assyriologists identify the most likely incorrect lemmata. The lowest scores of 0 and 1 are assigned to OOV words containing logograms and to syllabic spellings. The score of 2 is assigned to highly ambiguous in-vocabulary words in previously unseen POS contexts. The second highest score of 3 is assigned to in-vocabulary words that show low ambiguity, and the highest score of 4 to lemmata that exist in previously seen POS contexts.

5 Evaluation

For evaluation, we train ten models for the first millennium BCE Babylonian texts from Oracc comprising ca. 500,000 words in total. We use a 90/10/10 train/dev/test split and estimate the model's accuracy against two baseline models by using 10-fold cross-validation.[5] Our first **baseline** model is a dictionary-based lemmatizer and POS-tagger that labels the word forms in our test set with their most common lemmata and POS-tags seen in the training data. To measure the effect of our post-corrections, we use **TurkuNLP** without any post-correction scripts as the second baseline model. The results are presented in Table 1.

Model	Lemma	POS	Lemma+POS
Baseline	84.42 ±0.33	88.83 ±0.31	82.71 ±0.34
TurkuNLP	86.19 ±1.32	**97.32** ±0.10	85.31 ±1.31
BabyLemmatizer	**94.94** ±0.17	**97.32** ±0.10	**94.03** ±0.35

Table 1: Average accuracy (%) based on 10-fold cross-validation

In Table 2, we measure lemmatization accuracies in different confidence classes, as well as the proportion of lemmata that are assigned to each confidence class in our evaluation setting.

Confidence score	0	1	2	3	4
Accuracy	30.66%	56.71%	69.57%	96.25%	98.40%
Lemma-%	0.86%	3.87%	0.49%	52.10%	42.67%

Table 2: Confidence score distribution.

[5] In this experiment we use the default network architectures for training TurkuNLP's lemmatizer and tagger

5.1 Manual Evaluation

To test our lemmatizer, we apply it to a sub-corpus of Achemenet comprising 107,778 words. This is an Akkadian corpus with a different genre and time period distribution than our previous test sets. We use a model trained with the same Oracc data and train/dev/test split as in our evaluation setting described above, with an added glossary of Akkadian personal names from Prosobab (Waerzeggers and Groß et al., 2019). We then generate glossaries of the most common words that were assigned with the two lowest confidence classes and manually correct lemmata and POS-tags for word forms in the glossary file that have a frequency of >3 (for class 0) and >5 (for class 1) in the data. There were 315 unique corrected word forms, comprising 3.87% of the unique word forms covering 4.77% (5,037) of the 107,778 words in the sub-corpus.

To measure the accuracy of the lemmatizer and the effect of our manual corrections, we randomly select texts from our lemmatization results amounting to ca. 1,000 tokens for manual evaluation. We first evaluate the initial lemmatization without any manual corrections to the glossaries as a baseline. Then we apply our corrections to the lemmatization results in two ways: first, as a part of our *master glossary* of unambiguous lemmata (used in step 1 of post-processing), and second, by adding our manual corrections to the training data for TurkuNLP to see how much the system can learn from the corrections. The training data is added by first lemmatizing the text with a corrected master glossary and then replacing all words with the lowest two confidence scores with underscores to prevent the neural network from learning likely erroneous lemmata. Results are shown in Table 3.

	Lemma	**POS**	**Lemma+POS**
Baseline	93.0%	94.6%	90.2%
Glossary Override	96.2%	96.0%	93.8%
Retrained NN	**96.6%**	**96.1%**	**94.5%**

Table 3: Improvement in accuracy after corrections.

As can be seen from Table 3, our Lemma+POS labeling accuracy improves 4.3% when manually correcting only 3.87% of the unique word forms. The final results can be considered satisfactory for our current needs, which is to make the corpus searchable in Korp.

6 Conclusions

We presented a hybrid lemmatizer and POS-tagger for Akkadian, and demonstrated an increase of ca. 10% in Lemma+POS labeling accuracy compared with our baseline models. We also tested the lemmatizer on a previously unlemmatized Akkadian corpus with a different chronological and genre distribution than our training data. This test demonstrated that the system can reach a Lemma+POS labeling accuracy close to 95% after minor manual corrections.

References

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

Dominique Charpin. 2014. Ressources assyriologiques sur internet. In *Bibliotheca Orientalis, 71(3-4).*, October.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142, Brussels, Belgium, October. Association for Computational Linguistics.

Massimo Maiocchi. 2019. Thoughts on ancient textual sources in their current digital embodiments. In S. Valentini and G. Guarducci, editors, *Between Syria and the Highlands: Studies in Honor of Giorgio Buccellati and Marilyn Kelly-Buccellati*, pages 262–268. CAMNES.

Steve Tinney. 2019. *L2: How it Works, http://oracc.org/doc/help/lemmatising/howl2works.*

Wolfram von Soden. 1995. *Grundriss der akkadischen Grammatik (3rd edition)*. Pontifical Biblical Institute, Rome.

Caroline Waerzeggers and Melanie Groß et al. 2019. *Prosobab: Prosopography of Babylonia (c. 620-330 BCE), https://prosobab.leidenuniv.nl*.

WebLicht-Batch – A Web-Based Interface for Batch Processing Large Input with the WebLicht Workflow Engine

Claus Zinn
Department of Linguistics
University of Tuebingen, Germany
claus.zinn@uni-tuebingen.de

Ben Campbell
Department of Linguistics
University of Tuebingen, Germany
ben.campbell@uni-tuebingen.de

Abstract

WebLicht is a workflow engine that gives researchers access to a well-inhabited space of natural language processing tools that can be combined into tool chains to perform complex natural language analyses. In this paper, we present WebLicht-Batch, a web-based interface to WebLicht's chainer back-end. WebLicht-Batch helps users to automatically feed large input data, or input data of multiple files into WebLicht. It disassembles large input into smaller, more digestible sizes, feeds the resulting parts into WebLicht's pipelining and execution engine, and then assembles the results of such processing into files that preserve the usual input-output dichotomy.

1 Introduction

WebLicht is a web-based application that allows users to easily create and execute tool chains for linguistic analysis. No software must be downloaded or installed as all computation is delegated to tools that WebLicht knows about and interacts with on users' behalf (Hinrichs et al., 2010).

For a couple of reasons, WebLicht has a size limit on the data that users can upload for processing. First and foremost, WebLicht must take into account the analysis capabilities of the services it gives access to. While some services can cope with a large amount of data, others struggle with much less data to process. Second, WebLicht needs to keep the computation time of the services connected to WebLicht within a reasonable limit, and network-related socket timeouts need to be avoided, if possible.

In this paper, we present WebLicht-Batch, a browser-based service built upon the WebLicht backend that helps users to invoke WebLicht with large input. Our work also supports users that need to process a set of text files at once. Rather than submitting them manually to WebLicht, users can upload them as a collection archive so that WebLicht-Batch can process the collection one by one. Both usage scenarios are intertwined with each other in cases where a collection of files contains one or more large files.

2 Background

WebLicht is an execution environment for natural language processing pipelines. It uses a service-oriented architecture (SOA), where web services can be combined into processing chains. Chains are executed via sequential HTTP POST requests to services on the chain; here, the output of service n is the input to service $n+1$ in the chain. Most services in WebLicht use Text Corpus Format (TCF)[1] as their input and output, and each service usually adds one or more annotation layer(s) to the result file.

The WebLicht GUI provides users with a web-based interface to upload their data and get it processed by their NLP pipeline of choice. In WebLicht's Easy Mode, users can choose among pre-defined processing chains that match often-used linguistic pipelines. In Advanced Mode, users are supported to build *permissable* tool chains to customize or finetune the processing for the intricacies of the task at hand.

WebLicht as a Service (WaaS) is a REST service that executes WebLicht chains. Unlike the WebLicht web application, WaaS does not require a browser, and hence prevents browser-specific issues from arising such as file size upload limits. Also, it does not impose on users to perform the rather mundane

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

task of actioning a GUI to get processing started. With WaaS, users can run chains from their UNIX shell, scripts, or programs. Once users have defined a chain in the WebLicht browser interface, they can download the chain, and then they execute a HTTP POST request with the multipart/form-data encoding to invoke WaaS with the chain in question and the input data.

Note, however, that WaaS is not always the solution to process a single large file, or a collection of smaller files. First, there are some services in the WebLicht tool space that cannot handle large files *per se*. Once they fail on large input, the entire processing chain fails and no output is returned to users. In this case, users will need to manually split the input into smaller entities, get them processed one by one, and assemble the individual results into a compound entity. Also, some users are not comfortable mechanising such enterprise with a program script.

3 WebLicht-Batch

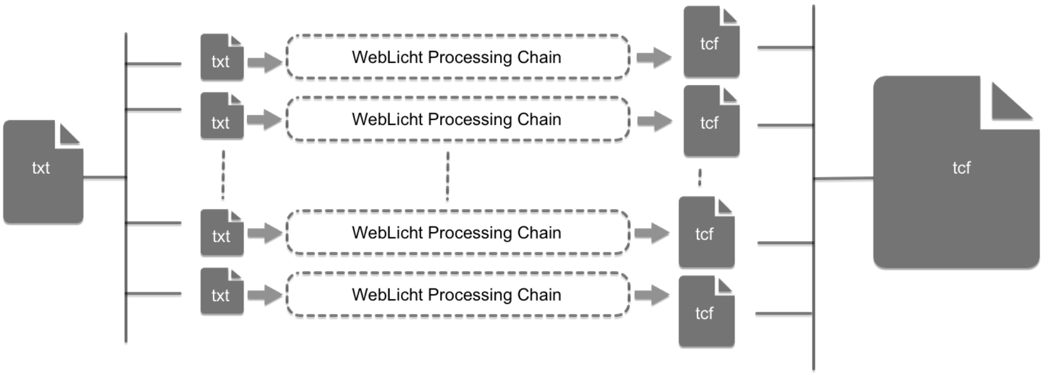

txt
txt
WebLicht Processing Chain
tcf
txt
WebLicht Processing Chain
tcf
txt
WebLicht Processing Chain
tcf
txt
WebLicht Processing Chain
tcf
tcf

Figure 1: WebLicht-Batch – Central Idea

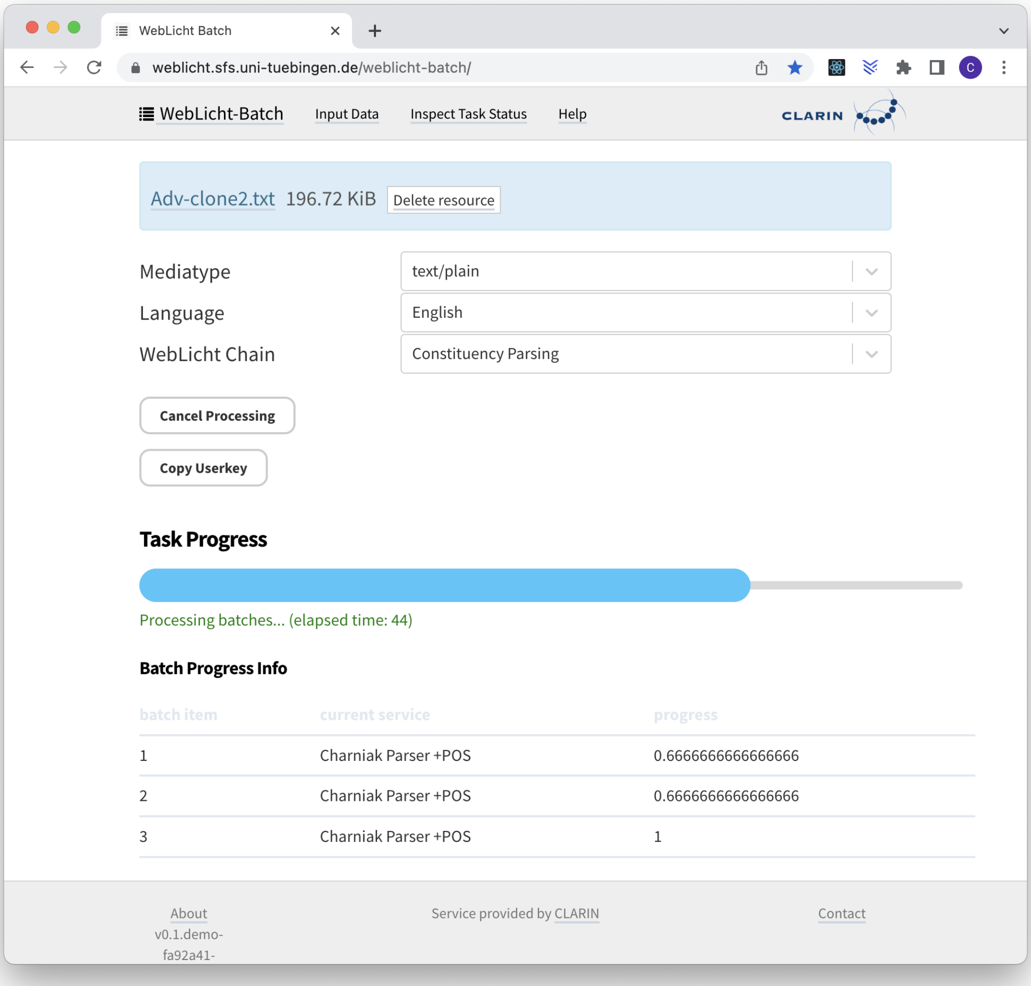
Fig. 1 depicts the central idea of WebLicht-Batch. A large plain text input file is split into multiple smaller files at sentence boundaries. Each individual file is then sent to WebLicht's pipelining and execution engine that processes the file with the NLP pipeline chosen by the user. The result of processing each file is captured in TCF format; they are then assembled to form a compound TCF-based result file. When users submit a zip file to WebLicht-Batch, each file in the archive is processed is the same manner. In addition, a zip file is constructed that contains the results of processing the individual files.

WebLicht-Batch makes use of WebLicht's pipelining and execution engine and provides, in addition, an API to upload a file (in plain text format, or zip format), to upload a chain, to start (or cancel) the batch process, to get processing information, and to retrieve the result file. The front-end of WebLicht-Batch makes use of this API and guides users through the overall process. Weblicht-Batch, hence, joins the WebLicht GUI and WebLicht As a Service as a third "user" of the Pipeling and Execution Engine.

WebLicht-Batch Front-End. Fig. 2 depicts the main GUI of WebLicht-Batch. Here, users can upload a single file, which can either be a plain text file, or a collection thereof being archived in zip format. Users then select the language of the text file(s) they want to process and also the processing chain they would like to run on the file(s). WebLicht-Batch gives access to all easy-chains offered by the WebLicht GUI, but users can also upload their own processing chain.[2] When users then press the "Start Processing" button, the batch-processing is started. Also, a user-specific key ("userkey") is generated that users are encouraged to copy to their clipboard. The user-key allows users to inspect the task status at a later time, even if they closed the browser tab in the mean time.

The figure also depicts the task progress for a plain text file that we have given to WebLicht-Batch. The text file has a size of just over 200 kilobytes and was split into a batch of three files. For each of the

[2] Processing chains are represented in an XML-based format. Users are advised to define, test, and download them using WebLicht's Advanced Mode.



Adv-clone2.txt 196.72 KiB

Mediatype: text/plain
Language: English
WebLicht Chain: Constituency Parsing

Task Progress

Processing batches... (elapsed time: 44)

Batch Progress Info

batch item	current service	progress
1	Charniak Parser +POS	0.6666666666666666
2	Charniak Parser +POS	0.6666666666666666
3	Charniak Parser +POS	1

Figure 2: The WebLicht-Batch GUI

three files, a table lists the progress, including the service that is currently run for each batch item.

WebLicht-Batch Back-End. The first step is the splitting of the original input text file into 100KB chunks, a size that most WebLicht services are comfortable with. This is somewhat of a ”chicken and egg” problem since, in order to split the file, it is necessary to use NLP tools which can perform this splitting, but which we do not want to feed too large of a file into, which requires the files to be split before sending them into the file splitter. In order to resolve this issue, we make use of the UDPipe tokeniser and sentence splitter (Straka and Straková, 2017) and feed in 100KB sized chunks – this size was chosen for the sake of convenience, as it is the same as the chunk size we use to perform batch processing. Splitting the file at 100KB results in text files which are split at arbitrary points in the final sentence. We start by sending the first chunk into the UDPipe tokeniser and sentence splitter, and assume that the last sentence of the output is incomplete, and then remove this sentence from the output of the first chunk, then add it to the beginning of the next chunk, which is then fed into UDPipe. This process is repeated until all chunks have been split into sentences. These chunks are then stored on the server to await further processing.

Next, we use the Weblicht chainer to process each chunk. At the time of this writing, the batch processor allows four chunks to be processed simultaneously as batches, which should allow a reasonable tradeoff between parallelism, and thus overall processing speed, and not overloading any of the services. Progress data, including which service of the chain is currently processing the chunk is constantly collected and sent to the frontend. If there is a failure in the processing of a chunk at any point, it is attempted to again run the chain on the chunk which failed. After three failures in a row, it is considered a failed batch and the entire task is considered to have failed.

If all batches succeed, the resulting TCF output files are then combined into one large TCF file. This is a complex process which involves manipulating the annotation layers for each TCF output file in order to ensure that the token ids for each token are correct for each annotation layer. If this combining is successful, a download link for the resulting file is sent to the frontend.

For a zip archive of plain text files, each file is processed as described above. The resulting TCF files are then packed into a zip file of which the download link is sent to the frontend. If the processing of any file in the archive fails, the entire processing is not considered to have failed. Rather, a list of files which have failed is kept and processing of the other files in the archive continues.

WebLicht-Batch has been integrated with CMDI Explorer (Arnold et al., 2020), a web-based tool that helps users exploring collections that are described with CMDI. In CMDI Explorer, users can select plain text files in the collection tree, request the generation of a zip file to bundle them, and send the archive to WebLicht-Batch for further processing. WebLicht-Batch has also been integrated with the Language Resource Switchboard (Zinn, 2018). When users upload a zip file to the Switchboard, WebLicht-Batch is shown as applicable tool. Once started, users are left to specify the common language of the text files and a WebLicht processing chain.

4 Discussion and Future Work

For large input, the WebLicht As a Service approach delegates the responsibility of file splitting at sentence boundaries and the combination of individual TCF files into a compound TCF to its users. Both file splitting and results' re-combination are non-trivial tasks that many users may not want to perform themselves. Those users will welcome WebLicht-Batch.

Apart from WaaS, we know of only one other application that addresses the processing of large data with WebLicht. But rather than splitting large input into more digestable chunks, it aimed at placing WebLicht services and the data they need to process into a shielded, high-performance environment – for big data (and also for sensitive data), it is better to move the tools to the data rather than having the data travel to the tools. In (Zinn et al., 2018), the EUDAT-based Generic Execution Framework (GEF) has been used to provide such environments. WebLicht services were installed in a so-called GEF environment with direct access to the data to be processed. A development version of WebLicht was built that had access to the GEF environment; and when users uploaded data to this version of WebLicht, the data was transferred to the location that also hosted the services.

The installation of GEF-ified services gives GEF maintainers the opportunity to preselect NLP services that can either cope with large data, or install many instances of the same service to handle many processing requests in parallel. While the installation of such purpose-built computing environments for the processing tasks at hand is costly, it helps minimising users' waiting times or processing errors. GEF itself was built using Docker software containerization technology, was seen as part the EUDAT Collaborative Data Infrastructure, but has never entered production mode; for more details, see `https://github.com/EUDAT-GEF/GEF`.

There are a number of issues that we would like to tackle in the future. Most services that are part of WebLicht's easy-chains are installed locally at institutional servers using Docker technology. For large input, we would like to investigate how to use Docker to spawn new workers of a given service on the fly giving a rising demand from WebLicht-Batch users. However, care must be taken to not overload individual services. A large WebLicht-Batch process could block regular WebLicht GUI users from getting their (smallish) input processed in time. Here, batch processing may want to postpone heavy processing to a point in time where Docker-based services are idle. Here, we may want to give users a scheduling option, where users are told estimated processing times depending on the time slots they choose.

From a practical perspective it is usually one service per chain that causes a bottleneck; this is usually a service offering constituency or dependency parsing, a rather complex process compared to tokenisation or part-of-speech-tagging. Here, we need to investigate whether complex processes should be given more CPU power and memory, or more workers by default, than simpler analyses.

In addition, more work is required to better understand the trade-off between the item sizes within a batch and the cost of splitting input into smaller chunks and the reassembling of individual results into a compound result. Also, the processing chain selected by WebLicht-Batch users should be taken into account. Chains without bottleneck services might profit from larger rather than smaller chunk splitting.

Most WebLicht services (usually not being part of easy-chains) are installed outside the control of WebLicht developers. Given the overall architecture of WebLicht and its few hundreds of services that are distributed over many different servers, batch design and task scheduling is all but trivial.

We invite all readers to test, play around, and use the service, which is available at `https://weblicht.sfs.uni-tuebingen.de/weblicht-batch`. Feedback is highly welcome.

References

Arnold, D., Campbell, B., Eckart, T., Fisseni, B., Trippel, T., and Zinn, C. 2020. CMDI Explorer. In Costanza Navarretta and Maria Eskevich, editors, *Selected Papers from the CLARIN Annual Conference 2020*, volume 180 of *Linköping Electronic Conference Proceedings*, pages 8–15. Linköping University Electronic Press.

Hinrichs, M., Zastrow, T., and Hinrichs, E. W. 2010. Weblicht: Web-based LRT services in a distributed escience infrastructure. In Nicoletta Calzolari et al., editor, *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta*. ELRA.

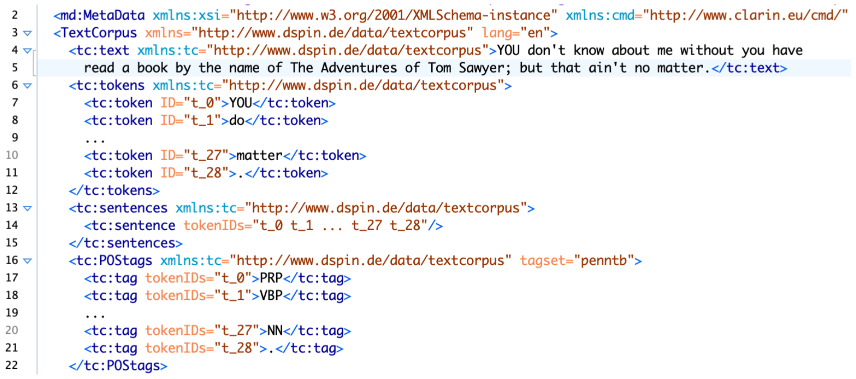
Straka, M. and Straková, J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zinn, C., Qui, W., Hinrichs, M., Dima, E., and Chernov, A. 2018. Handling Big Data and Sensitive Data Using EUDAT's Generic Execution Framework and the Weblicht Workflow Engine. In Nicoletta Calzolari et al., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan*. ELRA.

Zinn, C. 2018. The Language Resource Switchboard. *Computational Linguistics*, 44(4):631–639.

Appendix

Fig. 3 depicts a TCF fragment where the annotation layers for tokens, sentences, and part-of-speech tags have been added to the one-sentence input shown in the tc:text tag.



```
<md:MetaData xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:cmd="http://www.clarin.eu/cmd/"
<TextCorpus xmlns="http://www.dspin.de/data/textcorpus" lang="en">
  <tc:text xmlns:tc="http://www.dspin.de/data/textcorpus">YOU don't know about me without you have
    read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter.</tc:text>
  <tc:tokens xmlns:tc="http://www.dspin.de/data/textcorpus">
    <tc:token ID="t_0">YOU</tc:token>
    <tc:token ID="t_1">do</tc:token>
    ...
    <tc:token ID="t_27">matter</tc:token>
    <tc:token ID="t_28">.</tc:token>
  </tc:tokens>
  <tc:sentences xmlns:tc="http://www.dspin.de/data/textcorpus">
    <tc:sentence tokenIDs="t_0 t_1 ... t_27 t_28"/>
  </tc:sentences>
  <tc:POStags xmlns:tc="http://www.dspin.de/data/textcorpus" tagset="penntb">
    <tc:tag tokenIDs="t_0">PRP</tc:tag>
    <tc:tag tokenIDs="t_1">VBP</tc:tag>
    ...
    <tc:tag tokenIDs="t_27">NN</tc:tag>
    <tc:tag tokenIDs="t_28">.</tc:tag>
  </tc:POStags>
```

Figure 3: An abridged TCF example

The CLaDA-BG Dictionary Creation System: Specifics and Perspectives

Zhivko Angelov, Kiril Simov, Petya Osenova, Zara Kancheva
Institute of Information and Communication Technologies,
Bulgarian Academy of Sciences
Sofia, Bulgaria
angelov.zhivko@gmail.com, {kivs,petya,zara}@bultreebank.org

Abstract

The paper reports on the system for creating dictionaries within the CLaDA-BG infrastructure. At the heart of the system lies the BTB-Wordnet around which all other language resources are organized. These are various dictionaries, ontologies, corpora. The specific features and functionalities are outlined. Also, the rationale behind the construction of such a system are given.

1 Introduction

In this paper we present the main principles and perspectives behind the CLaDA-BG Dictionary Creation System — **CLaDA-BG-Dict**. The ultimate goal of its implementation is to support the compilation of new dictionaries by individuals or collaborators with respect to a certain task and through the usage of all the available resources within 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.

We aim to provide a system that supports the whole cycle of creating various types of dictionaries. At the heart of this system lies the Bulgarian BulTreeBank WordNet (BTB-WN) — (Osenova and Simov, 2018). It has been developed as an aggregator of semantic knowledge around which other dictionaries and sources of information (including grammatical, encyclopedic, etc.) have been organized in the form of a(n) (inter)linked knowledge network.

The motivation for the development of the CLaDA-BG-Dict refers to the need for: better control on the consistency in the creation of lexical language resources; user friendly and communicative collaborative environment; better connections among the available resources. Also in the light of open data we expect that there will be more lexicographical data available for reuse in future.This will facilitate the rapid creation of specialised lexicons as well as their publishing and individual usage.

The incentive for the design and implementation of CLaDA-BG-Dict system was the development of the BTB-WN. On the one hand, we were aware that there already exist software systems for the creation of other wordnets such as BulNet, GermaNet and Polish Wordnet (plWordNet). However, these systems reflect the approaches of the creators of these wordnets and they do not support all of the functions, needed for the work on BTB-WN, such as extension of the structure of lexical entries; mapping to other resources (inflectional lexicon, explanatory dictionaries, bilingual dictionaries, Wikipedia pages, etc,); concordance for selection of examples; ticketing system for identifying and handling errors of various types. On the other hand, the workflow on a contemporary Wordnet requires the addition of linguistic information beyond synsets, lemmas, definitions and relations. Such information includes: links to grammatical paradigms, valencies, links to Wikipedia, mappings to other wordnets, appropriate examples, etc.

2 Related Work

In our work we follow the approaches described in two existing wordnet editing systems — (Henrich and Hinrichs, 2010) and (Naskret et al., 2018). Similarly to (Henrich and Hinrichs, 2010) we needed to switch from a tool that supported only local editing where synsets were considered within a very limited context to a tool that supports editing of the wordnet data within a larger context. Comparably to both systems we switched from a file oriented presentation of the data to a centralized database used via web to support simultaneous work of a team of experts.

Here the following question might arise: Why to develop yet another system when there are already so many? We decided to implement our own system because we wanted to support also all language resources we already had incorporated within the current version of BTB-WN such as a spelling and grammar dictionary, explanatory dictionary, related corpora for providing adequate examples that show various sides of the respective meaning.

Our aim is to extend the current system further towards a full-fledged dictionary writing system. It is envisaged to provide the necessary environment and services for the compilation of ad-hoc and task-oriented dictionaries through the access to all the language data - starting from the existing dictionaries, corpora, encyclopedic knowledge, and others.

We are aware that many efforts have already been invested in dictionary creation systems from various points of view: formats and standards; approaches in the representation of the linguistic knowledge; implementation strategies, etc.

Here we mention only some of the related work. One of the most influential ongoing frameworks is ELEXIS[1]. After having performed an in-depth survey on the needs of lexicographers[2], the team behind ELEXIS (p. 62) envisages 'two complementary sets of tools will be provided: lexicographic workflow tools and crowdsourcing and gamification tools. The first will include a user-friendly open-source online dictionary writing system, with the aim to provide the central dictionary writing platform for new lexicography which also includes new possibilities of online collaboration. The other will provide tools for new techniques of dictionary creation, such as explicit or implicit crowdsourcing (gamification).' We plan to customize and adapt these tools to our framework as much as possible in our future work. Our current system also supports LMF formats but not in its full capacity. LMF files can be uploaded, edited and then saved outside the system. However, not all LMF metadata is visualized. There is no converter from the internal files into Lemon Standard and back[3]. At this point we rely only on the LMF-based converters. It should be noted that we aim to facilitate the work not only of the lexicographers but also of any other researcher groups and common users. Thus, we imagine helping teachers to compile a dictionary-minimum for their class; or a student to construct incrementally a learner lexicon of Bulgarian related to a language that they know, etc.

Within DARIAH-ERIC a standard for representation of dictionaries has been developed — TEI-Lex0[4]. This standard is supported also by ELEXIS.

Last but not least, the CLARIN-ERIC Lexica Resource Family overviews 89 lexica, most of which are monolingual.[5]. They are of various types – inflectional, morphological, valency, multiword, stopwords, sentiment, etc. Thus, they are a good source for insights in adding more types of resources and more types of analyses into the system.

3 System Specifics and Functionalities

Initially CLaDA-BG-Dict was designed and implemented to support the verification and extension of the BTB-WN. The motivation for this was that the existing version of the BTB-WN was initiated in an XML format within the CLaRK System[6]. The XML format that was used was not quite standard because it had to reflect the incorporation of non-standardised data when creating the BTB-WN. This fact however exhibited some shortcomings. As mentioned above, the main problem with working in CLaRK system was that the lexicographers had only a local view (without an underlying database) over the existing Bulgarian synsets. For instance, it was not easy to see all the synsets in which a given lemma

[1] https://elex.is/

[2] https://elex.is/wp-content/uploads/2019/02/ELEXIS_D1_1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf

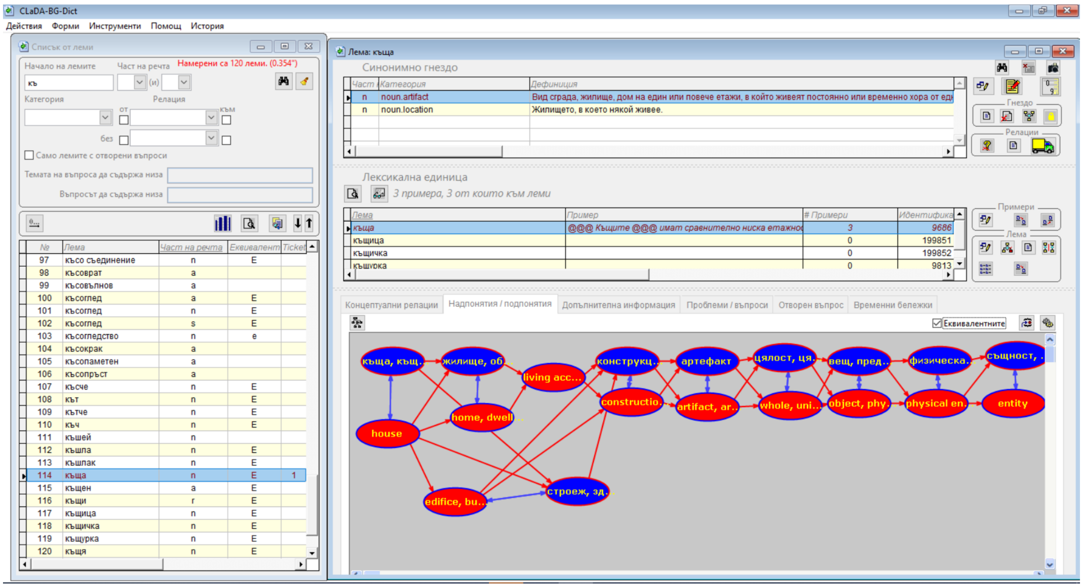
[3] https://www.w3.org/2019/09/lexicog/

[4] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[5] https://www.clarin.eu/resource-families/lexical-resources-lexica

[6] http://bultreebank.org/en/clark/

participates, because they could be in different XML files. One of the main design decisions was to support the mapping to the Open English WordNet (OEW) with the idea to enhance the multilingual applications and the transfer of information from the OEW to BTB-WN. In addition, we needed some system support for the better integration of BTB-WN with other language and knowledge resources for Bulgarian.



Начало на лемите: къ

Намерени са 120 леми. (0.354")

№	Лема	Част на речта	Еквивалент	Ticket
97	късо съединение	n	E	
98	късоврат	a		
99	късовълнов	a		
100	късоглед	a	E	
101	късоглед	n	E	
102	късоглед	s	E	
103	късогледство	n	e	
104	късокрак	a		
105	късопаметен	a		
106	късопръст	a		
107	късче	n	E	
108	кът	n	E	
109	кътче	n	E	
110	къч	n	E	
111	кышей	n		
112	къшла	n	E	
113	къшлак	n	E	
114	къща	n	E	1
115	къщен	a	E	
116	къщи	r	E	
117	къщица	n	E	
118	къщичка	n	E	
119	къщурка	n	E	
120	къщя	n	E	

Част	Категория	Дефиниция
n	noun.artifact	Вид сграда, жилище, дом на един или повече етажи, в който живеят постоянно или временно хора от едн
n	noun.location	Жилището, в което някой живее.

3 примера, 3 от които към леми

Лема	Пример	# Примери	Идентифика
къща	@@@ Къщите @@@ имат сравнително ниска етажнос	3	9686
къщица		0	199851
къщичка		0	199852
къщурка		0	9813

Figure 1: A screenshot of the user interface of CLaDA-BG-Dict. It shows a search on several criteria of lemmas within the current BTB-WN (on the left); A synset editor (on the right) — for a selected lemma it shows all the synsets (upper part of the window), for the selected synset it shows the category, a definition and a list of synonyms. For each lemma there are assigned examples as well as mappings to inflectional paradigms. At the bottom of the window a graphic representation of a noun hierarchy is given and also the mapping to English synsets.

The system is a client-server web-based editor using a thick client model. The thick client is installed on the user computer (desktop or laptop). The database is installed on a server and it is accessed online via the web. The data is stored in a relational database. Two people cannot work on one synset at the same time, but they can work subsequently. Also, the system stores in a log file each editing step, the name of the person who edited and at what time the edit was done. In this way we could track back states of the data and repair errors if necessary. The system reflects our approach towards the wordnet further development as well as towards the integration of various language resources within any dictionary compilation in order to reflect the lexicon-grammar interface in a better way. In Fig. 1 a screenshot is presented.

Regarding the BTB-WN, the system provides information about a selected lemma: its meanings (synsets) and associated examples; its internal relations as well as the mappings to the OEW; it also provides the ratio among the used relations. In case of equivalent synsets between BTB-WN and OEW, the Bulgarian synset inherits all the relations from the English synsets. In cases when these equivalent synsets have also mapped hypernyms or hyponyms, the respective relations enrich them as well. [7] After the inheritance of the relations the lexicographers (we do not use this term in its very strict professional way but rather conditionally, i.e. we include anyone who wants to compile a dictionary) have the possibility to change the relations — deleting some of them when not applicable or adding of new relations. The system also allows for definitions of new relations and some (limited) inference when a new relation

[7] Assuming that this is true for relations between concepts represented by the synsets.

is addition of its reverse relation, some transitive relations, and others. Also domain and range restrictions are taken into account.

In case of a problem, the lexicographer can put a ticket with a special marking that indicates the type of the problem. For example: Wrong definition, Spelling error, Wrong mapping to OEW, etc. Concerning editing, the lexicographer can delete a lemma in a synset; can add a new lemma or meaning; can delete and edit a definition or example; can delete a relation. The user interface is organized via several types of forms that display the corresponding information and provide mechanisms for editing of information.

The main form is literal-based (in other words, string-based). Depending on the part of speech and the meaning each literal forms one or more lemmas participating in different synsets. Thus, the editor form will show all synsets containing a given literal (string), independently from their part of speech. When selecting a literal from a list of literals already in the BTB-WN, a form is opened and it shows the set of all synsets in which the literal is included as a lexical unit. When selecting one of the synsets the user observes all the lexical units for the synset, the set of relations to other synsets, assigned examples for each of the lexical units in the synset. In addition to this there is a possibility to consult a graphical representation, related information in the available dictionaries in the system (currently four dictionaries are included, see next section), assigned tickets and additional comments. The system supports different kinds of users, which allows different tasks to be assigned to them. CLaDA-BG-Dict keeps a list with history of changes and allows observation of the performed actions by different users.

Regarding the integration of various language resources, the system provides a possibility to observe various dictionaries in alphabetical order of entries. But it also allows for searches in the in-house corpora by a chosen lemma. Moreover, it can show the grammatical paradigm of the lemma (in case of inflected part of speech).

4 Language Resources Supporting Dictionary Creation

As mentioned several times above, we aim at providing a system where the user will be able to exploit inside one system all the available dictionaries, corpora and services. Thus (s)he will be able not only to statically consult other dictionaries but also to search within corpora, to make concordances as well as to establish mappings.

In our view the necessary minimum of functionalities of such a system would include: an editor of lexical entries supporting different structures of interrelated elements; access to existing dictionaries and corpora; concordances and other materials. BTB-WN that has been fully developed in this system serves as a connector to other dictionaries and corpora through its synset and lemma information.

At the moment the integration of BTB-WN has been made with OEW, but a coverage comparison and linking between corresponding senses is planned to be done with several dictionaries such as Bulgarian Explanatory Dictionary and Bulgarian Inflectional Lexicon (currently partially done). In addition to dictionaries the system supports mapping to Wikipedia via the inclusion of Wikipedia article URI to the corresponding synset. For the moment the interpretation is equivalent concepts, but more elaborated set of relations are necessary.[8]

Each of the included language resources inherits its structure defined in some standard (with some modification if necessary). For example, for a given lexicon included in the system the structure of the lexical entry will be presented in TEI Lex0[9]. Some other lexicon might be presented in Lemon or LMF. Thus, the user will be able to refer to the structure of the various lexicons, to extract parts from the lexicon entries, to combine elements from different types of lexicons. This will allow easy ways of reusing the available data. Another benefit of the system is that it can keep track on the provenance of the work threads.

When the new dictionary (lexicon) is shared within the system itself together with the relations to other resources the result will be a valuable resource not only for supporting the future dictionary creation, but also for automatic processing.

[8] We plan to adopt an already existing schema like SKOS, LEMON, etc.

[9] `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`

5 Conclusions and Future Work

CLaDA-BG-Dict is an editor, which could be used both for creating lexical databases like wordnets as well as traditional types of dictionaries. But what is more - it provides possibilities of linking the available data in many ways depending on the goal.

CLaDA-BG-Dict has already been successfully used for editing of more than 19 000 synsets that were created at earlier stages in an XML format, and for the addition of around 13 000 synsets together with appropriate examples.

It thus provides quick access to various types of linguistic resources and information – dictionaries, corpora, concordance, etc. The resources are accessible in the system, so any kind of checks could be performed by the user in the same environment with one click.

Our vision for future is to enhance replicability and re-usage of dictionary compilation as much as possible. In this way we believe that the work of both groups is facilitated and enriched - of dictionary creators and of dictionary users.

Last but not least, in its beta-version now the system uses its own format for uploading corpora and other digitally-born or digitized dictionaries. However, it is planned to conform to the common standards such as TEI, TEI LEX0, Lemon, etc. All the participating resources will be made available through the CLaDA-BG repository and dedicated web services.

Acknowledgements

This work was 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 DO01-301/17.12.21.

References

Henrich, V. and Hinrichs, E. 2010. *GernEdiT - The GermaNet Editing Tool.* Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

Osenova, P., Simov, K. 2018. The data-driven Bulgarian WordNet: BTBWN. Cognitive Studies – Études cognitives, 2018(18) (2018) https://doi.org/10.11649/cs.1713

Naskret, T. and Dziob, A. and Piasecki, M. and Saedi, Ch. and Branco, A. 2018. *WordnetLoom – a Multilingual Wordnet Editing System Focused on Graph-based Presentation.* Proceedings of the 9th Global Wordnet Conference. pp. 190–199

A Lightweight NLP Workflow Engine for CLARIN-BE

Adriaan Lemmens
Centre for Computational Linguistics
KU Leuven, Belgium
adriaan@ccl.kuleuven.be

Vincent Vandeghinste
Instituut voor de Nederlandse Taal
Leiden, the Netherlands
vincent.vandeghinste@ivdnt.org

Abstract

This paper presents our work in progress on building a flexible workflow engine. The architecture of the engine is based on the message queue programming model and is implemented in Python. NLP tools are exposed as remotely executable tasks and wrapped in standard abstractions. The main contribution is to provide a unofirm API for defining various kinds of tasks and workflows.

The use case of the library is building a text analytics platform for digital humanities, providing a usage-friendly interface to predefined workflows.

1 Introduction

CLARIN's infrastructure boasts a wealth of tools for Natural Language Processing (NLP) that, while individually capable, are non-trivial to combine. This is because each tool makes its own assumptions about the design of its interface (whether this be a GUI or an API endpoint), the data formats it supports, and whether it processes documents individually or in batches. Users who require the combined functionality of multiple tools face the daunting task of verifying whether tools' input and output formats are compatible and, moreover, of manually coordinating the execution of these tools. To remedy this, we are developing a new text analytics dashboard application (Section 3) that bundles existing tools into curated but configurable automated annotation workflows, which are invoked from within a cohesive, user-friendly environment. The responsibility of interoperating disparate tools is no longer the responsibility of individual users, but now moves to the application backend.

Tackling the complexity of flexible tool interoperability has led us to design and implement our own custom NLP workflow engine.[1] Like alternatives, our engine assumes an agreed-upon vocabulary of abstractions for defining multi-step data processing logic (aka 'workflows'), but focuses in particular on the needs of text processing. The engine is implemented as a Python library with a carefully designed API intended for use by non-core contributors, e.g. computational linguists at other research groups, giving them a clear path for registering new tools and workflows. The design of its architecture is informed by the finite deployment and maintenance resources of research infrastructure. Its core consists of a minimum of moving parts, most of them third-party open-source; when idling, its footprint is minimal; it is easy to set up and interact with locally.

2 Related Work

There is no shortage of open-source workflow engines for data pipelining in general. Popular options include Apache Airflow,[2] Prefect,[3] and Argo.[4] These are each backed by a committed community and/or commercial entity and have well-tested APIs. A first downside is that, in order of mention, they rely partly

[1] The engine is developed separately from the dashboard application, but its design is informed by the latter's requirements.

[2] https://airflow.apache.org/

[3] https://www.prefect.io/

[4] https://argoproj.github.io/

to fully on Kubernetes[5] for deployment.[6] We prefer to avoid a dependency on Kubernetes at this stage to avoid added complexity.[7] A second downside to these options is that, being generic workflow engines, they offer more flexibility than we need, yet their learning curves are steep. Our ideal workflow API is more limited, so that non-core maintainers need only learn a few abstractions to get started. Limitations here also ensure that contributed tasks and workflows behave more predictably. Of course, by rejecting these 'standard' options, we risk some degree of wheel reinvention, e.g. with respect to functionality such as error handling and logging as well as overall fault-tolerance. We also take on more responsibility in terms of maintaining good documentation and growing a community where fellow contributors can ask questions. The effort these last two entail ought not to be underestimated.

There appears to be less community energy for workflow engines that focus on a specific domain, such as NLP. Most effort is concentrated at the level of individual libraries, e.g. spaCy[8] (Honnibal et al., 2020) and Stanza[9] (Qi et al., 2020), where pipeline components are assumed to implement the library's abstractions.

Within CLARIN, the most established counterpart to our work is the *WebLicht* environment (Hinrichs et al., 2010), which has continued to receive updates over the years. For a user, WebLicht is a web application for chaining together linguistic annotation services and applying these chains to uploaded texts. This functionality is similar but not identical to that of our text analytics dashboard application (Section 3). At the architectural level, each linguistic annotation tool in WebLicht's inventory is exposed as a stateless HTTP web service.[10] WebLicht services process data within the life cycle of a single HTTP request, meaning requests must be conservative about both the number of input items (i.e. documents) they batch together as well as total item size. Annotation services can be implemented in any programming language in principle; however, the language of choice appears to be Java. Services are registered not in one central database, but across separate repositories hosted at CLARIN-D research centers. They are described by CMDI[11]-formatted metadata, which allows for them to be aggragated automatically into a central inventory. The accompanying metadata also includes information about input data and output data schemas, allowing annotation pipelines to be validated for correctness, an idea that we have borrowed. In Section 4 we indicate a number of important design differences between our system and WebLicht.

A second CLARIN project with similar goals is the distributed parallel NLP processing architecture built by CLARIN-PL (Walkowiak, 2017). This powers *ws.clarin-pl.eu* and implements the LMPN language for planning processing pipelines and orchestrating services. It has been in use for many years and according to on-page statistics it has processed millions of requests by services and applications built on top of it. Like Weblicht, it is an open source project[12]. From the code base, we glean several architectural similarities with our system, which we indicate in Section 4.

3 Use Case: Text Analytics Platform for Digital Humanists

The primary client of our workflow engine is a text analytics platform for digital humanists, which we are simultaneously developing. This future contribution to CLARIN-BE's infrastructure is essentially a dashboard application that lets non-technical users manage corpora, execute carefully curated analysis workflows, visualize results, and export annotated datasets to one of several standard formats.

An important difference between *Manatee* and an environment such as *WebLicht* (see Section 2) is that users are not expected to define workflows themselves, that is, starting at the level of individual tools and chaining them together.[13] Instead, users choose from an assortment of workflow 'presets' and configure them to suit their needs. From the point of view of our workflow engine, these presets are full-fledged

[5] https://kubernetes.io/

[6] Granted, Airflow can be deployed without Kubernetes, but such a setup nevertheless entails many moving parts.

[7] While Kubernetes is the de facto standard for container orchestration, affording great power and flexibility, it has a steep learning curve, its passive compute requirements are higher, and maintaining a cluster is extra work.

[8] https://spacy.io/

[9] https://stanfordnlp.github.io/stanza/

[10] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/WebLicht_in_Detail

[11] http://www.clarin.eu/cmdi

[12] https://gitlab.clarin-pl.eu/nlpworkers

[13] WebLicht does offer predefined processing chains in its so-called 'Easy Mode', albeit it less prominently.

workflows behind the scenes. Our main aim in taking this approach is to shield user interfaces from complexity.

4 Workflow Engine Architecture

The architecture of our workflow engine is based on the message queue[14] (MQ) programming model. Decoupled services communicate among each other by posting and pulling messages to/from a centralized queue data structure. In our particular case, the messages themselves contain instructions to execute a unit of work (or *task*) in the consuming service (*worker*). The effect is a form of remote procedure calling (RPC).

Our system is implemented as a Python library. Each worker service in our system is a containerized Python process that exposes one or more tools (CLARIN-affiliated or other) as remotely executable tasks. Tools are wrapped in standard abstractions defined by the library. Both clients and workers are aware of these abstractions, simplifying interoparability, as components are able to introspect each other's metadata and all components respect certain fixed contracts of behavior. The main contribution of the library is to provide a uniform API for defining various kinds of tasks as well as workflows. The API has been designed carefully so as to be easy-to-use for computational linguists in an academic setting, who by and large write code at an intermediate level and are relatively inexperienced with software engineering in general. A few important principles guiding the API's design are 1) that users can rely on their IDE to spot mistakes (for the same reason, we prefer metadata to live in-code rather than in, e.g. ad hoc YAML schema), 2) that testing locally is easy, and 3) that no knowledge of the system as a whole is required.

While CLARIN-PL's engine (Section 2) implements its own logic for interacting with a message queue, our framework delegates instead to the very established third-party Python library *Celery*[15]. As a result, we get at no extra cost functionality such as task planning, logging, fault recovery, best practices, and monitoring (not to mention extensive documentation) The most complex and crucial component of our system is thus in a sense offloaded to the open source community.

Another third-party library that we rely on is JinaAI's *DocArray*[16], which provides data structures for modeling documents and batching them together. All workflow steps are defined to operate on (possibly large) batches of documents, making techniques such as clustering and sentence alignment possible. Memory usage is not a concern, since DocArray provides on-disk versions of its document collection abstraction with random read and write access. The only data exchanged by workflow steps are pointers to the location of these on-disk datasets in shared storage.

Our engine requires Python 3.6 or greater[17]. Note, however, that tools implemented in other languages or in older versions can still be wired up via the Python API by wrapping them in subprocess calls.

Architecturally, our engine assumes one or more worker services and any number of clients, all of them communicating in the vocabulary of abstractions defined by our Python library. Additional essential engine components are 1) shared storage, where workflow step outputs are cached and 2) a message broker (in our case, Redis). Contributors of workflows and tasks do not define worker services directly. Instead, they instantiate a *Toolset* object and register tasks and workflows with it. Typically, a single Toolset should correspond to an independent code repository. Each Toolset must specify a Dockerfile specification of the required environment. Toolsets connect to the rest of the engine through a separate *Worker* abstraction (which non-core contributors need not know of). There are several kinds of tasks, or *Operator*s, that can be registered with a Toolset, each with their own function signature. The most common is the *Processor*, which takes a batch of documents, assigning one or more annotations to each document. Tasks are registered by decorating ordinary Python functions with a decorator method of the Toolset object, e.g. `@toolset.processor()`. These Python functions wrap CLARIN tools in their body. These tools are run on demand only. Toolset authors are recommended to keep actions that impact memory usage (e.g. loading a large model) inside the function body and rely on garbage collection to

[14] https://en.wikipedia.org/wiki/Message_queue

[15] https://docs.celeryq.dev/en/stable/

[16] https://github.com/jina-ai/docarray/

[17] Because we rely on Python type hints.

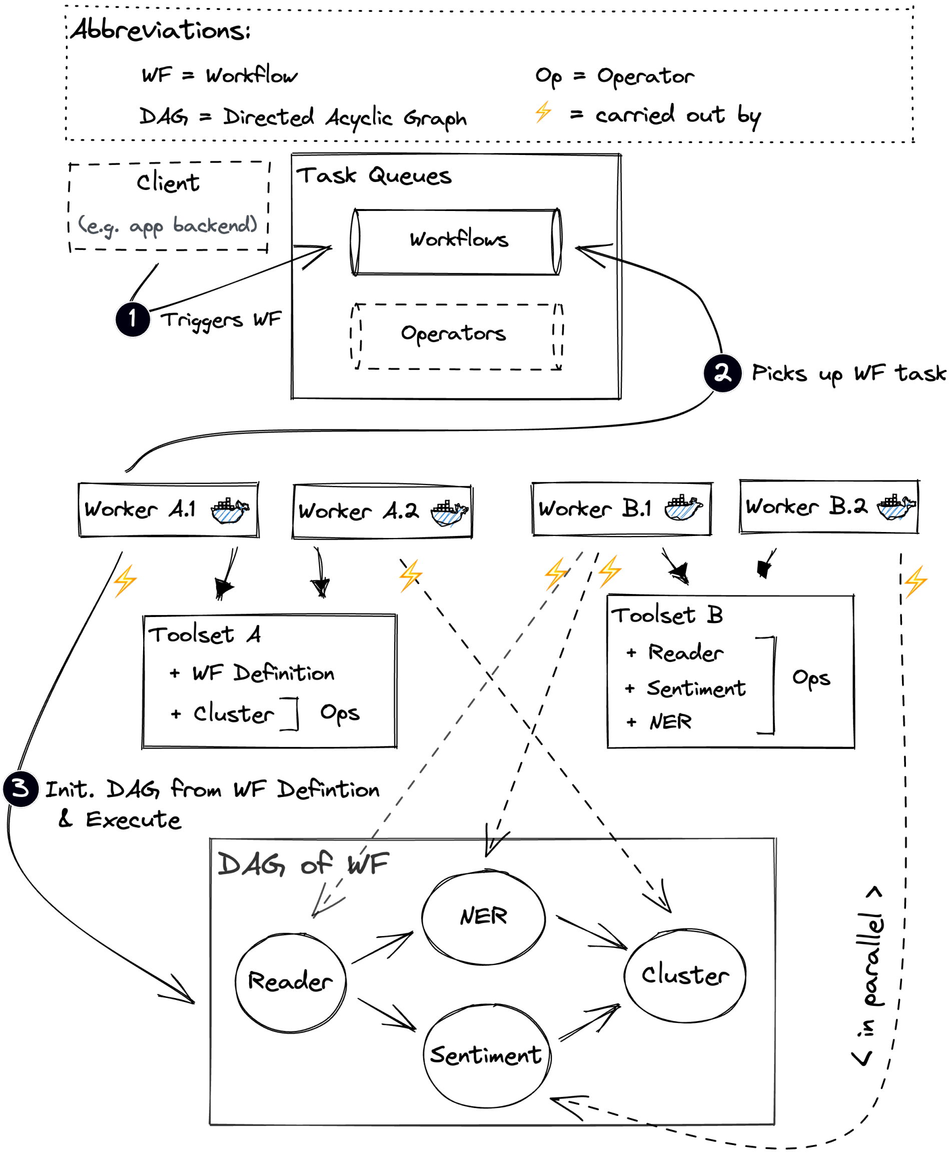


Figure 1: Life cycle of a workflow

release memory. It is the additional responsibility of the author of the task-wrapped function to implement the logic of transforming corpus data so as to match the input and output formats of the tool in question. Task metadata is provided through the Python API. Among other things, metadata is used by the engine to validate whether certain tasks are mutually compatible.

Workflow definitions are registered with Toolsets in much the same way as are tasks, i.e. with a corresponding Python decorator. Our workflow definitions can describe not only sequential data flow logic – aka ‘chains’/‘pipelines’ – but also more general *Directed Acyclic Graphs*[18] (DAG). The advantage of a DAG is that it allows for parallelization in the execution graph. A workflow defined in one Toolset can invoke Operators from a different Toolset, so long as both Toolsets are registered with the workflow engine. Moreover, because workflow exection logic is part of the core library shared by all worker services, there is no need for a central workflow orchestrator.

Our approach differs from WebLicht’s (Section 2) in that a) the workflow steps in our system can handle long-running batch jobs; b) the tools to which steps delegate run only when needed, c) our engine supports not only pipelines (serial step execution, aka ‘chains’ in WebLicht’s terminology), but also more advanced execution graphs with parallel branches; d) metadata and configuration are always specified through code. As for CLARIN-PL’s engine, which follows a similar task-as-message queue model to ours, we note that a) our workflow tasks are more strict and explict with respect to the types of input they accept; b) partly due to (a), our engine is less flexible; c) whereas CLARIN-PL’s engine separates functions cleanly across workers, our workers expose bundles of functionality (as Toolsets). CLARIN-PL is able to set precise resource usage limits on their workers and can scale horizontally more easily; by contrast, we need to take special care to ensure that workers don’t process too many messages at once, so as to avoid memory contention within the same worker (and possible deadlocks as a result).

A schematic overview of the life cycle of a workflow invocation is shown in Figure 1.

5 Early Results & Project Status

Our workflow engine is still undergoing active development and the API is far from stable. Nevertheless, early (and tentative) results suggest that our approach has merit. We are able to invoke a large number of workflows at once: once the task queue becomes saturated, remaining invocations are simply queued up. This is acceptable for our needs, since there are no system components that require low latency, e.g. a pending HTTP request somewhere, and users of our dashboard application (*Manatee*) will be conditioned through UX to expect longer processing times.

An alpha release of *Manatee* is planned for December 2022, with a beta release planned for Spring 2023. Around the same time, all code repositories will be made publicly available. After a v1.0 release mid-2023, incremental improvements will continue into 2024.

6 Discussion

At least two of our design choices warrant a critical look. First, our choice for a batch processing model, where a batch corresponds to an entire corpus, has the consequence of making it harder to predict the duration of a workflow step. The larger the input data, the longer a step will take to finish processing it. Without restrictions on corpus size, this can lead to a single workflow job holding a worker service hostage, causing slowdowns for other jobs. Our current solution to this is to deploy replicas of each worker service, insofar as our compute budget permits.

Second, because each worker service bundles multiple operators of varying memory requirements, and supposing that we permit workers to spawn multiple operator processes in parallel, it is hard to predict how much memory a worker will use at any given time. In order to keep the memory usage of the system as a whole reasonable, we set resource limits on the containers hosting the worker services based on observed behavior. An alternative solution, which is unappealing at the moment, is to differentiate between ‘heavy’ and ‘light’ operators and associate the heavier sort with a separate queue that lets fewer tasks through at once.

[18] https://en.wikipedia.org/wiki/Directed_acyclic_graph

References

Erhard W. 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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Tomasz Walkowiak. 2017. Language processing modelling notation - orchestration of NLP microservices. In Wojciech Zamojski, Jacek Mazurkiewicz, Jaroslaw Sugier, Tomasz Walkowiak, and Janusz Kacprzyk, editors, *Advances in Dependability Engineering of Complex Systems - Proceedings of the 12th International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, July 2-6, 2017, Brunów, Poland*, volume 582 of *Advances in Intelligent Systems and Computing*, pages 464–473. Springer.

Natural Language Processing for Literary Studies: Graph Literary Exploration Machine (GoLEM)

Agnieszka Karlińska
Institute of Literary Research,
Polish Academy of Sciences, Poland
agnieszka.karliska@ibl.waw.pl

Wiktor Walentynowicz
Department of Computational Intelligence, Wroclaw University of Science and Technology, Poland
wiktor.walentynowicz@pwr.edu.pl

Jan Wieczorek
Department of Computational Intelligence, Wroclaw University of Science and Technology, Poland
jan.wieczorek@pwr.edu.pl

Maciej Maryl
Institute of Literary Research,
Polish Academy of Sciences, Poland
maciej.maryl@ibl.waw.pl

Abstract

This paper presents a design of a web-based application for the analysis and visualisation of relations between terms, named entities, and topics. The goal of this project is to create, in close cooperation between the literary research community and the IT professionals, a comprehensive workflow tailored to the specificity of literary studies at large, and, the current debates and trends in humanities research. The application brings together the already existing tools offered by CLARIN-PL and the resources and tools developed at Dariah.lab. It consists of three components: the named entity relationship analysis component, the terminology extraction component and the topic modelling component. The whole system is not only based on an automatic operation of the components but is constructed to allow user intervention in individual sub-processes of the entire processing pipeline. A strong emphasis is put on the use of metadata of the analysed texts (e.g., for filtering and grouping documents) and the visualisation of results.

1 Introduction

GoLEM (Graph Literary Exploration Machine) is a new web-based application for advanced analysis and visualisation of synchronic and diachronic relations between terms, named entities, and topics, tailored to the needs of literary scholars. In our presentation, we will introduce the main theoretical and methodological considerations behind GoLEM, its architecture and possible applications. We will also demonstrate how GoLEM stands out from other web-based text analysis tools.

2 Design Considerations

GoLEM is the next-generation service built upon the experience with LEM (Maryl et al., 2018) which is part of the CLARIN-PL infrastructure and one of the primary research environments for Polish textual scholars. GoLEM was designed as a solution to the problem of scattered NLP-based literary research tools in Polish and their inadequacy for both the specificity of literary studies at large and, for the current debates and trends in humanities research. We assume that the development of data exploration and analysis tools requires close cooperation between the research community and IT professionals. Moreover, such cooperation should be driven by theoretical challenges and hypotheses created in a disciplinary context, so the tools would be attuned to the actual challenges of the user community. With few exceptions, the currently available web-based applications for analysing large textual collections were not designed with the actual needs of literary scholars in mind but rather were intended for the widest possible range of users from different disciplines. This also applies to LEM, which was initially addressed mainly to literary scholars, but has built its user base beyond this discipline and is developed in the framework of projects in various SSH disciplines. GoLEM is designed as a less versatile tool, but better suited to specific uses, material, and research questions.

GoLEM brings together the state-of-the-art NLP solutions offered by CLARIN-PL and the resources (dictionaries, corpora, and language models) and tools developed at Dariah.lab, a research infrastructure for digital humanities. The tools are combined in a comprehensive workflow and adapted to process a specific type of text: scientific papers. In the first version of GoLEM, we will focus on Polish literary papers. In the future, other SSH disciplines will be addressed. The solutions deployed for GoLEM are not language agnostic, so expanding the system into languages other than Polish is not a trivial matter. The GoLem platform is, however, designed in such a way that expansion with modules in other languages is possible through international cooperation with other teams/projects. The platform coordinating such development could be CLARIN ERIC or the national CLARIN/DARIAH/CLARIAH consortia. Unlike other CLARIN-PL tools, a strong emphasis in GoLEM is put on the visualisation of results, e.g. as graphs, time series, maps or scatter plots.

Another distinctive feature of GoLEM is the close relation between uploaded texts and their metadata. Most NLP-based text analysis tools make only limited use of bibliographic data. In GoLEM it is possible to add metadata to documents in bulk, filter texts by metadata and group texts based on metadata or keyword searches in order to perform comparative analyses of subcorpora. GoLEM provides the possibility to work on ready-made corpora, in particular on the Corpus of Polish Literary Studies Discourse, which includes texts from the last 200 years selected according to strict methodological criteria, or the user-made corpora stored in different text formats.

Currently, the design phase is completed and GoLEM is in production. The testing phase of the individual components will start in mid-2023. The release of all deliverables of the Dariah.lab, including GoLEM, will take place in early 2024.

3 Scheme of the System

The data processing module on which the GoLEM application is based has three main constituent components (Fig. 1). These are the named entity relationship analysis component, the terminology extraction component and the topic modelling component.

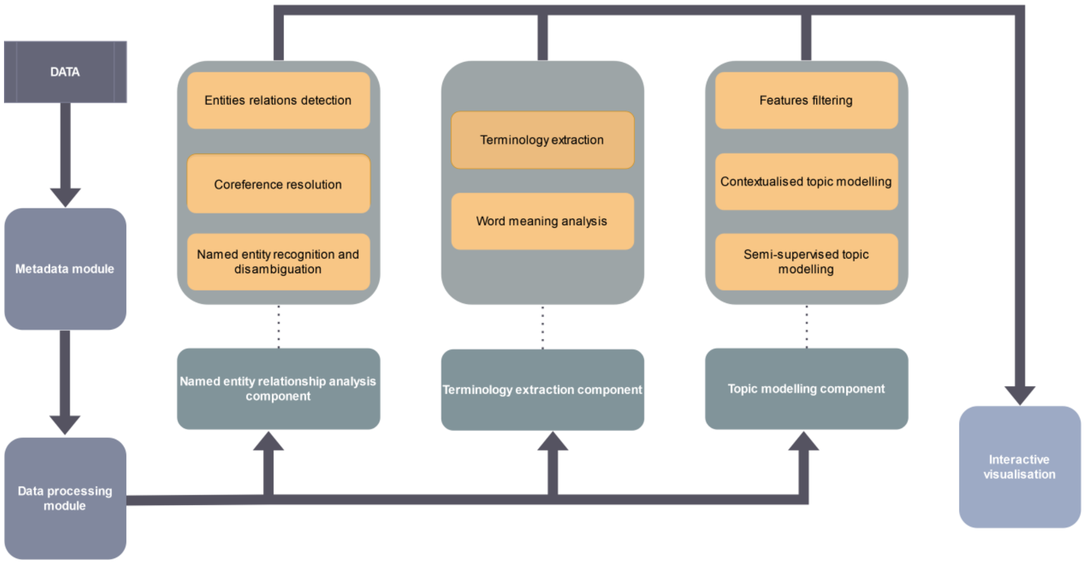

DATA
Metadata module
Data processing module
Entities relations detection
Coreference resolution
Named entity recognition and disambiguation
Named entity relationship analysis component
Terminology extraction
Word meaning analysis
Terminology extraction component
Features filtering
Contextualised topic modelling
Semi-supervised topic modelling
Topic modelling component
Interactive visualisation

Figure 1. GoLEM architecture.

The named entity relationship analysis component performs the tasks of named entity extraction, named entity disambiguation with available knowledge bases and linking entities with relations. Named entity extraction is based on a system using pre-trained language models in Transformer technology, i.e. BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2020), XLM-RoBERTa (Conneau et al., 2020), trained for the task of named entity extraction on available corpora. The use of such technology allows the system to be independent of lexical resources and allows the tool to be easily extended between different languages – in the future, GoLEM could be used to analyse more than one language. GoLEM localises and extracts not only standard named entity categories, but also categories specific to literary studies: names of authors, researchers, translators, literary groups, literary works, journals, cultural institutions, cultural events, fictional characters and imaginary places. The pipeline includes person name clustering and coreference resolution. The output of the named entity extraction module is the input for the entity relationship detection module. The module uses a neural network-based classification system to determine the occurrence of relationships of a given type between entities. Relations (e.g., cooperation, contribution, affiliation, location) are detected by taking into account the contexts in which they are located. The scope from which entities in a relation may come (e.g., a sentence, a paragraph, a document) is defined at the user level. The last module of this component is the entity disambiguation module. It uses the connection to resources in the form of knowledge bases and contention information to unambiguously indicate what or who is being referred to in the text. This module is also based on neural models, using the context in which the entities occur.

The terminology extraction component uses pattern search methods to extract terminologies described in knowledge bases, including a comprehensive dictionary of literary terms mapped onto plWordNet. The terminology extraction process consists not only of indicating occurrences but also of performing statistics at the corpus and document level. It also allows, with the help of a model of distributional semantics, to study changes in the meaning of terms over time.

The topic modelling component allows topic identification, key phrase extraction and stylometric comparison. It is based on clustering methods, classification using machine learning and statistical methods for comparing vector representations. In addition to the LDA technique, semi-supervised topic modelling and contextual topic modelling that uses pre-trained representations of language to support topic modelling (Bianchi et al., 2021; Bianchi et al., 2021a) are implemented. The component allows for filtering of features used in the process, identification of characteristics and presentation of results. The module is capable of incorporating user's commands in order to suggest analysis directions or to set obligatory rules. The data obtained from this component can be used to construct resources such as frequency dictionaries, semantic networks or categorisation of documents.

The whole system is not only based on an automatic operation of the components but is constructed to allow user’s intervention into individual sub-processes of the entire processing pipeline. This is an important element in the application to working with scientific papers, as it allows the user to use their domain knowledge to improve their performance and avoid error propagation from previous modules. In this way, GoLEM is implicitly designed as a decision support system that can be used as an end-to-end system if the user gives up interference.

4 Use Cases

GoLEM takes into account different levels of user experience: basic (users interested in processing a small number of texts and qualitative exploration of results), advanced (users interested in the processing of larger corpora and further work in other tools, e.g. Gephi and Neo4j), expert (users interested in using preprocessed data to test innovative methods and workflows).

The design process for GoLEM took as its starting point the real needs of users deriving from specific research questions formulated by researchers of literary texts or the underlying theoretical frameworks. GoLEM allows for the identification of constellations of terms and concepts and the comparison and visualisation of their flows over time, between literary and academic communities, within a single discipline, between different disciplines or from literary theory to artistic practice. It enables interdisciplinary research in the study of “travelling” concepts in the humanities (Bal, 2002), historical semantics, cultural analysis, philosophy of science or sociology of knowledge.

GoLEM will allow for research designs aimed at addressing the following issues:

- showing how the reception of a given author has evolved
- reconstructing and analysing the comparative terminology networks of selected scholars

- grasping differences in the understanding of key concepts between representatives of different fields of literary studies
- analysing the semantic field of a selected literary term
- detecting words with usage change across corpora
- tracing how the topics of selected researchers' papers have changed over time
- analysing the links between authors based on the topics discussed in their papers
- identifying academic writing styles and argumentation building in the humanities

The three modules are designed as complementary. To provide robust answers to more complex research questions, it is advisable to employ more than one component. For example, for the reconstruction of literary thought collectives understood as communities bound by the exchange of ideas or intellectual interaction (Fleck, 1979), named entity relationship component and topic modelling component can be used simultaneously: the first one will allow for identifying the co-occurrence of researchers' names in cooperation contextes, and the second one – for exploring the authors' networks based on the topics covered in their papers by linking texts to topics and then replacing them with their authors and visualising the results in the graph form. The combination of the two components will enable to determine which author, in whose texts a given topic was recognised, was most often referred to by other authors addressing the same issue.

5 Conclusion

The operational idea behind GoLEM is to convince scholars that distant reading – the analysis of multiple texts at once using statistical methods (Moretti, 2013) – is not an approach that competes with close reading but rather facilitates literary interpretation by providing new insights into analysed texts, GoLEM's role is to facilitate the use of digital methods of text analysis and enable new answers to research questions that are well established in literary discourse.

References

Bal, M. 2002. *Travelling Concepts in the Humanities. A Rough Guide*. University of Toronto Press, Toronto.

Bianchi, F., Terragni, S., and Hovy, D. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Volume 2: Short Papers. Association for Computational Linguistics: 759–766.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., and Fersini, E. 2021a. Cross-lingual Contextualized Topic Models with Zero-shot Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics: 1676–1683.

Conneau, A., Khandelwal, K., et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: 8440–8451.

Devlin, J., Chang, M., Lee K., and Toutanova, K. 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.* Association for Computational Linguistics, Minneapolis, Minnesota: 4171–4186.

Fleck, L. 1979. *Genesis and Development of a Scientific Fact*. University of Chicago Press, Chicago.

Liu, Y., Ott, M., et al. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ICLR 2020. https://openreview.net/forum?id=SyxS0T4tvS.

Maryl, M., Piasecki, M., and Walkowiak, T. 2018. Literary Exploration Machine: A Web-Based Application for Textual Scholars. In *Selected Papers from the CLARIN Annual Conference 2017*. Linköping University Electronic Press, Linköpings universitet: 128–144.

Moretti, F. *Distant Reading*. 2013. Verso, London.

Supporting Ancient Historical Linguistics and Cultural Studies with EpiLexO

Valeria Quochi[1], **Andrea Bellandi**[1], **Michele Mallia**[1], **Alessandro Tommasi**[2], **Cesare Zavattari**[1]
[1]Institute for Computational Linguistics, National Research Council, Pisa, Italy
name.surname@ilc.cnr.it
[2]Department of Computer Science, University of Pisa, Italy
alleytom@gmail.com

Abstract

This contribution presents a system of independent software components meant to support the creation of ecosystems of interrelated language data (i.e. lexica linked to textual testimonies, concepts, metadata, bibliographic references, and other external lexical resources) according to the current state-of-the-art representational models for the semantic web. The system is implemented as a set of autonomous servers exposing Restful APIs that in principle can serve different front-end applications and use cases. In this work they serve the EpiLexO GUI application designed and geared to support scholars of ancient languages of fragmentary attestation in their studies. The development of both the back-ends and the front-end is still work-in progress, but a first version is ready for use.

1 Motivation

The practice of editing electronic dictionaries is nowadays long-standing and several models and tools already exist for encoding lexical data and supporting scholars in their manual creation or revision work. Most available technology however is designed for the representation of contemporary, modern or classical highly attested languages (e.g. Latin or Old Greek). Very little exists for *Restsprachen* ‘languages of fragmentary attestation’, i.e. dead languages that have reached us only through limited written testimonies, mostly in epigraphic form (inscriptions, stamps, coins), and whose reconstruction is substantially partial both in terms of grammar and lexicon (Rigobianco, 2022). These languages pose peculiar challenges to standard lexical models: as it is often impossible to have a complete attestation of a declension or paradigm, for instance, lemmatisation cannot be appropriately operated in many cases, making an alternative representation preferable. Etymological information is also fundamental, albeit often partial, and uncertainty is pervasive at various levels of representation. Intense work was carried out in recent years within the lexical modelling community (think in particular at Ontolex (Cimiano et al., 2020), TEI Lex0 (Romary and Tasovac, 2018) and LMF (Francopoulo and Monte, 2013; Romary et al., 2019)) which brought about useful extensions and adaptations of existing lexical models. The time seems thus ripe to bring together such advancements and the important digital turn in the historical linguistics and epigraphy communities for developing tools that would facilitate the creation and exploration of (interlinked) digital materials for ancient languages, and make them compatible with the current Semantic Web.

Most existing lexicon creation tools, such as VocBench (Stellato et al., 2020), do not fully support the special requirements mentioned above, do not yet provide easy linking to (digitised) attestations, and are designed as full-stack applications that may not be easily adapted to different usages. By the same token, existing corpus management tools, such as TEITOK (Janssen, 2016), or EFES (The Digital Classicist Wiki, 2021) generally lack direct integration with lexical data and other external resources.

In this present work we attempt to devise a technological solution that, while satisfying the needs of a specific project[1], may be flexible enough to serve different use cases. The focus of this contribution is thus on a system of independent software components that currently serve a web GUI application, EpiLexO,

[1] The project *Languages and Cultures of Ancient Italy: Historical Linguistics and Digital Models* (ItAnt hereafter) studies

developed with the aim to support historical linguists and more generally scholars of ancient languages and cultures in the creation of interlinked linguistic resources according to well-accepted representational models. For reasons of space this abstract is deliberately focused on technical aspects; details on the current ItAnt use case that may be of more interest from the user perspective are given in Quochi et al. (2022).

1.1 Relations to CLARIN

EpiLexO can be considered as an upgraded and augmented version of the CLARIN-IT existing tool LexO[2], as it extends it with functionalities for linking lexical information to various other external resources, including their textual attestations. Project ItAnt is a CLARIN-IT sponsored project[3] with the mandate to provide data and tools to the CLARIN-IT infrastructure for better serving its community of historical linguists and digital humanists. The project will thus share within CLARIN several (interlinked) datasets: a multilingual lexicon for the languages of ancient Italy (Oscan, Faliscan, Venetic and possibly Celtic), a related bibliographic data set, and an EpiDoc XML corpus of new editions of inscriptions from the cultures of ancient Italy. Data and code is and will be made as FAIR as possible. Although devised within a specific funded project, the system has been designed to be general and the EpiLexO application may be useful to other projects dealing with similar tasks and data. For reasons of space we just mention for instance the *Cretan institutional inscriptions* project that recently deposited and published its data in CLARIN-IT[4] (Vagionakis, 2021; Vagionakis et al., 2021).

Although still work-in-progress, a first version of the system is complete and currently being tested[5].

2 The EpiLexO System of Back-ends

The whole system is realised as a Service-Oriented Architecture with strong front end-back end separation of concerns in such a way that its services are reusable in different contexts and applications. The implementation of each component, i.e. servers and clients, is done independently, and the services are made as general as possible. The server side is composed of two main back-ends: the LexO-server and the CASH-server, which manage lexicons and (annotated) texts respectively. They expose APIs based on the HTTP protocol and exchange data in JSON format.

The **Lex**icon and **O**ntology server (**LexO-server** henceforth) deals with lexicons and ontologies and manages the linguistic and conceptual dimensions independently, but intimately linked. The linguistic model is based on *OntoLex-Lemon*[6] (Cimiano et al., 2020), while the conceptual part is currently based on the Simple Knowledge Organization System (SKOS). The usage of a standard model helps harmonising scholars' work and increasing the quality of the produced data. LexO-server re-engineers as REST services all the functionalities of the full-stack LexO-lite version described in Bellandi (2021) and adds new features, such as a federated query on specific external SPARQL endpoints and a plug-in to Zotero for linking bibliographic references, accessed via its APIs. The code is open source[7] and the APIs are documented and testable online[8].

The **C**orpus, **A**nnotation and **S**earc**H** server (**CASH** henceforth) deals with the management of (text) documents, annotations and metadata. Document management is exposed so as to allow for a file-system-like handling of documents. Annotations are represented as complex objects that minimally anchor a main annotation value to the specific text location by means of spans of character; an annotation may

the forms of linguistic variability before Romanisation by investigating the cultures of ancient Italy on the basis of their linguistic documentation, and bringing together methods and practices from traditional linguistics, philology, and computational lexicography.

[2] LexO-lite is an existing, full-stack, collaborative web editor for easily building and managing lexical and terminological resources for the Semantic Web (Bellandi, 2021; Bellandi et al., 2019). PID: http://hdl.handle.net/20.500.11752/ILC-95.

[3] https://www.clarin-it.it/en/content/launch-prin-project-interest-clarin-it

[4] http://hdl.handle.net/20.500.11752/OPEN-548

[5] At the time of writing the whole system can be installed following the instructions provided here https://github.com/DigItAnt/epilexo-stack and tested online in a demo version https://lari2.ilc.cnr.it/LexO-angular/lexicon

[6] Ontolex-Lemon is the *de facto* standard for representing linguistic data in the Semantic Web https://www.w3.org/2016/05/ontolex/

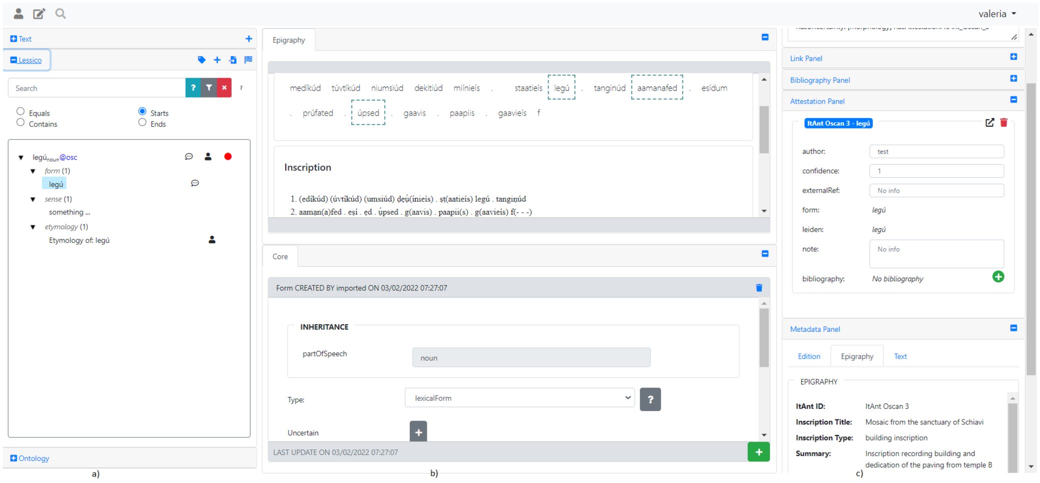
[7] https://github.com/DigItAnt/LexO-backend

[8] https://lari2.ilc.cnr.it/LexO-backend-itant/

also contain additional features (e.g. annotation layer, author, confidence) and can be enriched with an arbitrary number of attribute-value features. Values can also consist simply of URIs, thus allowing for linking operations. Metadata are specified at document level and can be typed. At present, CASH can ingest documents either in the form of plain text files, or TEI EpiDoc files XML files, which thus contain annotations and metadata, as in the case of our native usage scenario. CASH APIs are documented and testable online[9] and the code is open source[10]. Finally, AAI and user management are handled via an independent Keycloak server instance[11].

3 The EpiLexO Front-end GUI

This section describes how the servers described above are used in the ItAnt project to build a web application dedicated to the creation and editing of lexical resources for ancient fragmentary languages linked to their 'testimonies' (i.e. transcriptions of epigraphic texts), to related bibliography, to contextual metadata and to other relevant independent resources available on SPARQL end-points such as the LiLa Knowledge Base (Mambrini et al., 2020). The web services exposed by the servers are invoked by the **EpiLexO GUI**[12], a front-end designed as a single-page web application made up of several components. Each component provides a set of functionalities that allow for creating and interlinking a lexical resource with various the various external datasets. EpiLexO typical target users are historical linguists whose work lies very much at the crossroad between linguistics and philology and for which the availability of (textual) testimonies is a necessary precondition for carrying out their linguistic analyses.



Inscription

1. (edíkúd) (úvtíkúd) (umsiúd) deṷ(inieís) . st(aatieís) legú . tanginúd
2. aaman(a)fed . esí . ed . úpsed . g(aavis) . paapii(s) . g(aavieís) f(- - -)

Form CREATED BY imported ON 03/02/2022 07:27:07

partOfSpeech: noun

Type: lexicalForm

Uncertain

LAST UPDATE ON 03/02/2022 07:27:07

author: test
confidence: 1
externalRef: No info
form: *legú*
leiden: *legú*
note: No info
bibliography: *No bibliography*

ItAnt ID: ItAnt Oscan 3
Inscription Title: Mosaic from the sanctuary of Schiavi
Inscription Type: building inscription
Summary: Inscription recording building and dedication of the paving from temple B

Figure 1: The EpiLexO GUI - (a) Resources tree; (b) Editing panels; (c) Contextual information

The platform GUI, shown in Fig: 1), is divided into three main vertical sections. The left column (a)) contains the navigation trees for the main resources: corpus, lexicon, ontology[13]. The corpus tree is based on the CASH-server APIs and is organised like a common file system: in this panel the user can e.g. create, delete, rename or move folders and files, and import new files to be ingested. The lexicon tree is populated with the Lexo-server APIs, it allows to browse the lexicon content and provides some filtering options. It is organised according to the key ItAnt lexical classes: Lexical Entry, Form, Sense and Etymology, which dynamically correspond to dedicated editing areas in the central part of the interface.

[9] https://lari2.ilc.cnr.it/cash

[10] https://github.com/DigItAnt/CASH-server; the code is written in Java with the Spring Boot framework; the persistence layer is based on MySQL DBMS.

[11] https://www.keycloak.org/; Keycloak has been chosen also in light of the future integration with the CLARIN AAI federation as it supports the SAML standard.

[12] EpiLexO is developed in Angular (https://angular.io/) and the code is open source (https://github.com/DigItAnt/Epilexo

[13] The LexO-server implements services for importing an external ontology and linking its elements to lexical items, a feature which is currently not exploited in project ItAnt.

From this panel the user can also perform some high-level actions: e.g. add and manage languages, create new lexical entries, mark an Entry as ready to be revised or completed, etc.

The right column (c)) contains panels dedicated to various kinds of “accessory information” that are either ingested from the input files or added within the interface: the Link panel is for encoding links to external datasets: for example,this is where external reference to the LiLa Lemma Bank (Passarotti et al., 2020) for Latin cognates may be encoded in the ItAnt lexicon. The Bibliography panel is for adding bibliographic references to lexical items via Zotero and enrich them with additional information; the Attestation panel is where text linking can be visualised and where is can be enriched with bibliographic references, and other optional information; the Metadata panel displays metadata related to the inscriptions and their editions, as encoded in the original input files (the EpiDoc documents in the ItAnt case). The content of these right panels is contextual, i.e. dynamically dependent on the items selected in the left or central part of the interface. For instance, the metadata panel will show the metadata related to the current inscription selected in the corpus navigation tree, whereas the bibliography panel will show the bibliographic reference linked to the lexical item selected in the lexicon tree.

The central part, b) in Fig. 1, is the main working area devoted to the editing operations. It consists of two horizontal sections: the lower part contains the Lexicon Editor, the upper part the text Linker. The Lexicon Editor is pivotal to the whole platform, it is modular and contextually adaptive, i.e. it shows editing web-forms on the basis of the item selected in the lexicon tree. Editing is dynamic: changes are directly recorded and registered in the LexO-server back-end. As EpiLexO presently makes use only of a subset of the Ontolex model, it is currently possible to encode information for lexical entries, forms with attestations, senses and etymologies. The possibility of encoding lexical concepts will be added soon,

Finally, the Linker displays the text of the inscription contained in the current document and allows for the linking of text spans to items in the lexicon, by invoking the services of the CASH-server[14].

Linking can be done first by selecting an entire token (by clicking on it), a sub-token (by dragging inside a token to select the desired portion, e.g. a prefix), or a list of tokens (e.g. for multi-words), then searching and selecting the desired corresponding morphological form in the lexicon within the dedicated pop-up window[15]. A wizard guides the lexicographer in case the form or the entire entry needs to be created in the lexicon. Text-lexicon links are treated as special annotations, i.e. as Attestations, and refer directly to the text spans, with the token id being an attribute of the annotation. The act of establishing a link between a text portion and a lexical form practically corresponds to creating an Attestation for the given Form.

Although the EpiLexO system was conceived and developed within the ItAnt project (see §1) and thus tailored on the requirements of *Restsprachen* of ancient Italy, the web services exposed by the two main servers are indeed quite general. The granularity of the APIs, in principle, make them capable of serving different applications and use cases that involve e.g. (LLOD) dictionary creation / update / post-editing, text linking, entity linking, among others. EpiLexO might be useful for instance in supporting post-editing of automatically acquired lexicons and text linking. In a completely different scenario, some of the LexO web services are already successfully invoked to add query expansion functionalities to a system that exploits a contemporary Italian lexicon (Giovannetti et al., 2021).

Future work includes the completion of the editing platform with more import/export, the addition of multi-level and cross-resource search functions, and the development of the “fruition” mode for consultation and visualisation of the whole resource ecosystem.

Acknowledgements

This work was carried out in the context of the “Languages and Cultures of Ancient Italy. Historical Linguistics and Digital Models” project (PRIN 2017XJLE8J), funded by the Italian Ministry of University and Research. The project also has the support of CLARIN-IT.

[14] Because the ItAnt EpiDoc corpus encodes word tokens, the Linker currently makes use of this information and displays the linkable text into visual segments.

[15] The back-end would also allow to select and link to any arbitrary text span, which may be more appropriate in a different scenario, but the GUI currently does not support this functionality.

References

Bellandi, A., Khan, A. F., and Monachini, M. 2019. Enhancing lexicography by means of the linked data paradigm: Lexofor clarin. In Simov, K. and Eskevich, M., editors, *Proceedings of the CLARIN 2019 Annual Conference*.

Bellandi, A. 2021. LexO: an open-source system for managing ontolex-lemon resources. *Language Resources and Evaluation*, 55.4:1093–1126.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. 2020. Modelling lexical resources as linked data. In *Linguistic Linked Data: Representation, Generation and Applications*, pages 45–59. Springer International Publishing, Cham.

Francopoulo, G. and Monte, G. 2013. Model description. In *LMF Lexical Markup Framework*, chapter 2, pages 19–40. John Wiley & Sons, Ltd.

Giovannetti, E., Albanesi, D., Bellandi, A., Marchi, S., Papini, M., and Sciolette, F. 2021. The role of a computational lexicon for query expansion in full-text search. In *CLiC-it*.

Janssen, M. 2016. Teitok: Text-faithful annotated corpora. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Mambrini, F., Cecchini, F. M., Franzini, G., Litta, E., Passarotti, M. C., and Ruffolo, P. 2020. LiLa: Linking Latin. risorse linguistiche per il latino nel semantic web. *Umanistica Digitale*, 4(8), Jan.

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F., Litta Modignani Picozzi, E., Moretti, G., Ruffolo, P., and Sprugnoli, R. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. studi e saggi linguistici. 58:177–212, 09.

Quochi, V., Bellandi, A., Khan, F., Mallia, M., Murano, F., Piccini, S., Rigobianco, L., Tommasi, A., and Zavattari, C. 2022. From inscriptions to lexica and back: A platform for editing and linking the languages of ancient italy. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 59–67, Marseille, France, June. European Language Resources Association.

Rigobianco, L. 2022. La linguistica delle lingue di attestazione frammentaria. In Meluzzi, C. and Nese, N., editors, *Metodi e prospettive della ricerca linguistica*, volume 29, pages 83–94. Ledizioni.

Romary, L. and Tasovac, T. 2018. TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources. In *TEI Conference and Members' Meeting*, Tokyo, Japan, September.

Romary, L., Khemakhem, M., Khan, A. F., Bowers, J., Calzolari, N., George, M., Pet, M., and Banski, P. 2019. LMF reloaded. *CoRR*, abs/1906.02136.

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., and Keizer, J. 2020. Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11(5):855–881.

The Digital Classicist Wiki. 2021. Epidoc front-end services — the digital classicist wiki. [Online; accessed 26-August-2022].

Vagionakis, I., Gratta, R. D., Boschetti, F., Baroni, P., Grosso, A. M. D., Mancinelli, T., and Monachini, M. 2021. Cretan institutional inscriptions meets clarin-it. In Monachini, M. and Eskevich, M., editors, *CLARIN Annual Conference Proceedings*.

Vagionakis, I. 2021. Cretan institutional inscriptions: A new epidoc database. *Journal of the Text Encoding Initiative [Online]*.

EU Data Governance Act: New Opportunities and New Challenges for CLARIN

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

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

Abstract

The Data Governance Act was proposed in late 2020 as part of the European Strategy for Data, adopted on 30 May 2022 (as Regulation 2022/868). It will enter into application on 24 September 2023. The Data governance Act is a major development in the legal framework affecting CLARIN and the whole language community. With its new rules on the re-use of data held by the public sector bodies and on the provision of data sharing services, its new limits on international transfers of non-personal data, and especially its encouragement of data altruism, the Data Governance Act creates new opportunities and new challenges for CLARIN ERIC. This abstracts briefly analyses the provisions of the Data Governance Act, and aims at initiating the debate on how they will impact CLARIN and the whole language community.

1 Introduction

The third decade of the 21st century has started with some of the most perturbing events in generations: the COVID-19 pandemic and the war in Ukraine, which – understandably so – overshadowed all other developments. Meanwhile, however, the European Union is dynamically modernizing its legal framework concerning digital data.

The GDPR and its application has shown that the EU is a global regulatory superpower. As one of the world's largest markets (in terms of population, but especially in terms of purchasing power), the EU has the power to influence the policies of manufacturers of goods and providers of services worldwide, by imposing standards on goods and services that can enter its market. With the so-called "Brussels effect" (Bradford, 2020), the GDPR has become the global standard for personal data protection. This phenomenon is now likely to extend to other types of digital data, markets, and services.

2 Towards the Data Governance Act

On 19 February 2020, just a couple of weeks before the first COVID lockdowns, the European Commission launched the European Strategy for Data (European Commission, 2020a). This was followed by a large stakeholders consultation (until 31 May 2020), in which 806 contributions were received, including 98 from academic/research institutions. A series of proposals for Regulations (labelled, in the Anglo-Saxon way, "Acts") were adopted based on this consultation, including:

- Data Governance Act (25 November 2020);
- Digital Services Act (15 December 2020);
- Digital Markets Act (15 December 2020);

- Artificial Intelligence Act (21 April 2021);
- Data Act (23 February 2022).

As of August 2022, the Digital Services Act and the Digital Markets Act in the final stages of their legislative process; the legislative processes regarding the Data Act and the Artificial Intelligence Act are ongoing; and the only one of these Acts that has already been adopted and published is the Data Governance Act (DGA). The Commission's proposal has, with few modifications, become the Regulation 2022/868 of 30 May 2022 on European data governance.

This abstract will briefly present the content of the DGA from the perspective of CLARIN ERIC and the EU language community as a whole.

3 The Content of the Data Governance Act

As expressly stated in the explanatory memorandum, the DGA was inspired by the FAIR principles. It has a variety of goals, which can be summarized as follows:

- To encourage wider re-use of personal data and data protected by IP, held by public sector bodies.

Under DGA, public sector bodies (such as public administrations – data held by cultural and educational establishments are expressly excluded from the scope of the DGA – Article 3(2)(c) of the DGA) are under a general obligation to make their datasets available for re-use, even if they contain personal data or data protected by third-party copyright (such data are currently excluded from the rules on re-use of public sector information, cf. Article 1(2)(c) and (h) of the Open Data Directive 2019/1024). This can be achieved e.g., by sharing pre-processed (anonymised) data (Article 5(3)(a)(i) of the DGA Proposal), or by limiting data sharing to secure processing environments (Article 5(3)(b) and (c) of the DGA). Where no other solution is available, the DGA imposes an obligation for public sector bodies to 'provide assistance' to potential re-users in seeking consent from data subjects, or permission from rightholders (Article 5(6) of the DGA). In order to assist public sector bodies in fulfilling these obligations (including to provide them with technical support, e.g. in the field of data anonymisation), Member States should create 'competent bodies' with adequate legal and technical capabilities and expertise (Article 7 of the DGA).

This creates many opportunities for CLARIN and the EU language community as a whole – first of all, a wealth of new data, including language data, will be made available for re-use, including for such purposes as developing language resources, and training language models. Secondly, the DGA will stimulate the development of anonymisation and pseudonymisation techniques and standards, as well as secure solutions for data sharing, which will also open new possibilities for language data. Thirdly, CLARIN's expertise in processing language data can potentially be interesting for the 'competent bodies', which will create new possibilities for liaising with other actors of data economy, and increase the visibility of CLARIN and its activities.

- To introduce a mechanism for supervising providers of data sharing services.

Under the DGA, the provision of 'data sharing services' is subject to notification (Articles 10 and 11). 'Data sharing services' are defined to include, among others (Article 10(a)):

intermediation services between data holders (...) and potential data users, including making available the technical or other means to enable such services; those services may include bilateral or multilateral exchanges of data or the creation of platforms or databases enabling the exchange or joint use of data, as well as the establishment of other specific infrastructure for the interconnection of data holders with data users.

The notification is to be made to a 'competent authority' (which are to be designated by every Member State), whose task is to monitor compliance with a set of obligations listed in Article 12 of the DGA, including, e.g. the prohibition of re-using data for other purposes than providing them to the users and the obligation to place data sharing services under a separate legal entity (Article 12(a)); the prohibition to use metadata collected from the provision of the service for other purposes than developing the service (Article 12(c)); the general obligation to keep the data in the format in which they were provided by the user (Article 12(d)) and the obligation to ensure continuity of service and access to the data in case of insolvency (Article 12(h)). Failing to meet these conditions may result in ‘dissuasive’ financial penalties and/or cessation of the provision of the service (Article 14(4)).

CLARIN ERIC (and/or CLARIN B-centres) undoubtedly fall within the definition of a data sharing service provider, and the obligations imposed by the DGA might be very difficult to comply with. However, the regulations mentioned above do not apply to “(…) not-for-profit entities insofar as their activities consist of seeking to collect data for objectives of general interest, made available by natural or legal persons on the basis of data altruism” (Article 15). It may therefore be necessary for CLARIN ERIC and for CLARIN B-centres to strictly limit their activity to collecting data made available on the basis of data altruism (see below).

- To promote ‘data altruism’

‘Data altruism’ is defined as ‘the consent by data subjects to process personal data pertaining to them, or permissions of other data holders to allow the use of their non-personal data without seeking a reward, for purposes of general interest, such as scientific research purposes or improving public services’ (Article 2(16) of the DGA).

DGA recognises ‘data altruism organisations’ as legal entities constituted to meet objectives of general interest (such as language research) which operate on a non-for profit basis independently from any for-profit entities (such as private research companies), and through a legally independent structure (separate from other activities). Data altruism organisations which meet these requirements are subject to registration with a competent authority (which are to be designated by each Member State); they are bound to respect transparency requirements *vis-à-vis* the data holders (Article 20(1) of the DGA; e.g., a full up-to-date list of entities granted access to the data should be provided together with the purpose of processing declared by each of those entities). Furthermore, these organisations are obliged to submit an annual activity report to the national competent authority (Article 20(2)); if applicable the report shall include a summary of results of the data uses allowed by the organisation. Article 22 of the DGA obliges the European Commission to adopt (in close cooperation with stakeholders) a rulebook laying down specific rules applicable to data altruism organisations, which shall include *inter alia* appropriate technical and security requirements for data storage, and recommendations on relevant interoperability standards.

The advantage of being granted the statute of a ‘data altruism organisation’ lie in the mechanism described in Article 25 and referred to as the ‘data altruism consent form’. The form will be adopted by the European Commission in a specific procedure and its exact content is impossible to predict; however, it seems that it will be possible for natural persons and legal entities to ‘donate’ their data to a ‘data altruism organisation’ in a way that ‘circumvents’ the some of the current regulations on data protection, and possibly also on intellectual property.

The mechanism of ‘data altruism consent’ could of course be extremely beneficial – it could potentially make most important legal hurdles in access to language data go away – but the status of a ‘data altruism organisation’ would require changes in how CLARIN centres operate. In particular, the data obtained through data altruism cannot be made available for re-use to everyone and for every purpose (i.e., under ‘open’ conditions), but only to researchers, possibly with an obligation to report back on the results. It remains unclear if a ‘data altruism organisation’ can also provide access to data obtained through other means than the ‘data altruism consent’, e.g. via open licenses or directly from the public domain. Finally, it is not clear whether ERICs are to be granted any form of special treatment with regards to this aspect of the DGA (such as automatic recognition as ‘data altruism organisations’).

In sum, for CLARIN ERIC, ‘data altruism’ is a central element of DGA, one which will certainly have to be discussed at the highest level. Unfortunately, as for today many elements of this framework remain unclear – but it is not too early to start a debate.

- To restrict transfer of non-personal data to third countries.

The transfer of personal data to third countries is already restricted by the GDPR (Articles 44-50). Many provisions of the DGA aim at restricting the transfer of non-personal data to third countries (i.e., outside the European Economic Area). This applies to non-personal data made available by public sector bodies, as well as by data altruism organisations (see above), which can only be transferred to third countries if certain obligations are met (see esp. Article 31 of the DGA).

This de-globalisation measure may turn out to be an obstacle for CLARIN's collaborations with countries from beyond the European Economic Area (EU, Iceland, Lichtenstein and Norway).

4 Conclusion

The Data Governance Act, after its entry into application in mid-2023, will be a major development in the legal framework affecting CLARIN and the whole language community. With its new rules on the re-use of data held by the public sector bodies and on the provision of data sharing services, its new limits on international transfers of non-personal data, and especially its encouragement of data altruism, the Data Governance Act will create new opportunities and new challenges for CLARIN ERIC.

References

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

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

European Commission. 2020b. *Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act)*. COM(2020) 767 final.

CLARIN Depositing Guidelines: State of Affairs and Proposals for Improvement

Jakob Lenardič
Department of Translation
University of Ljubljana, Slovenia
jakob.lenardic@ff.uni-lj.si

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

Abstract

The paper presents a review of the guidelines for depositing language resources in CLARIN B-centres. We discuss how the existing guidelines instruct depositors to document basic resource metadata such as size, annotation, and language. On the basis of our review, we propose a new set of guidelines to be adopted by CLARIN repositories for those metadata categories that are pivotal for the SSH community but are underrepresented in the existing guidelines.

1 Introduction

Distributed research infrastructures such as CLARIN are committed to providing high-quality metadata about the resources that they host in order to facilitate their use and reuse within the wider Social Sciences and Humanities (SSH) (De Smedt et al., 2018). However, there are differences in the rate of provision of the metadata and the quality of their documentation from one CLARIN repository to the other as well as between metadata categories within a single repository (McCrae et al., 2015; Cimiano et al., 2020). In terms of provision, for instance, it has been shown in the context of the CLARIN Resource Families Project (Lenardič and Fišer, 2020) that the metadata profiles of CLARIN-hosted language corpora lack information about linguistic annotation at a significantly higher rate than other types of structural metadata such as the size of the resource, even though annotation is typically the information external users looking for a corpus are most interested in judging by the requests sent to popular corpus-linguistics mailing lists such as Corpora-List.[1] Another prevailing issue, relevant from the perspective of digital source criticism within Digital Humanities (Koolen et al., 2019), is that there are quite a few deposited corpora that are insufficiently described from a qualitative perspective, in the sense that it is often unclear what the corpus contains relevant to the specific needs of its research domain.

For these reasons, we have conducted a review of the depositing guidelines of the service-providing CLARIN centres, looking at how they instruct depositors in documenting basic metadata such as size and resource annotation as well as writing up the resource description. On the basis of our review, we propose a new set of guidelines to be adopted by CLARIN repositories for those metadata categories that we believe to be the most pivotal for the SSH community but are underrepresented in the existing guidelines. The paper is structured as follows. Section 2 presents the survey. Section 3 presents the proposal for the new guidelines. Section 4 concludes the paper.

2 The Survey of the Depositing Guidelines

We have reviewed the depositing guidelines of the 23 repositories that are certified as CLARIN B-centres as of April 2022.[2] CLARIN B-centres are Service Providing Centres that offer the academic community sustainable access to resources, language technologies as well as knowledge in a way that is compliant with the FAIR data principles (Wilkinson et al., 2016). For metadata in a B-certified repository, the

[1] See https://office.clarin.eu/v/CE-2017-1098-CorporaList-report.pdf for a report on such user requests sent to Corpora-List.

[2] For a full list of CLARIN B-centres, see: https://www.clarin.eu/content/certified-b-centres

B-centre	Annotation	Size	Language	Name	FTD
ASV Leipzig	No	No	**Yes**	**Yes**	**Yes**
ARCHE	No	No	**Yes**	**Yes**	No
BAS	**Yes**	No	No	No	No
CELR	No	**Yes**	**Yes**	**Yes**	**Yes**
CLARIN:DK	No	No	No	No	No
CLARIN:PL	No	No	No	No	No
CLARINO	No	No	No	No	No
CLARIN.SI	No	**Yes**	**Yes**	**Yes**	**Yes**
TalkBank	No	No	No	No	No
EKU Tübingen	No	No	No	No	No
LINDAT	No	No	No	No	No
MIP	No	No	No	No	No
PORTULAN	No	No	No	No	No
SWE-CLARIN	No	No	No	No	No
ILC4CLARIN	No	No	No	No	No
FIN-CLARIN	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
ZIM	No	No	No	No	No

Table 1: The survey of the English instructions on documenting metadata across the CLARIN B-centres (FTD stands for "free-text description"). The sorting of the B-centres follows the one on the CLARIN ERIC webpage; see Footnote 2 for the hyperlink to the list of B-centres.

guidelines of the CLARIN centre assessment procedure require that they be both machine- and human-readable (Wittenburg et al., 2019, 7).[3]

For this survey, we have checked the depositing guidelines of the B-centres to see whether they provide specific instructions for the documentation of four types of metadata – *annotation*, *resource size*, *resource language*, and *resource name* – as well as instructions for providing the free-text description that accompanies each repository entry. We have limited ourselves on guidelines for linguistic corpora because the majority of the CMDI-metadata profiles used by the repositories are tailored to resources rather than tools (Odijk, 2019). The results of the survey are available in a Google Spreadsheet.[4] Out of the 23 B-certified repositories, 17 or 74% have guidelines that are available in English, while the guidelines for the remaining centres are either not in English (3 or 13%) or have not been identified (also 3). The guidelines are mostly presented as running text guiding the user through the depositing process,[5] or, less often, non-descriptively in a table indicating which of the metadata categories are obligatory or not.[6]

Table 1 briefly summarizes the surveyed inclusion of instructions on the aforementioned four metadata categories and the free-text description (abbr. FTD in the Table) in the English guidelines. What counts as included in the guidelines (labelled as "Yes" in the Table) is any kind of instruction on what a depositor needs to consider when providing the metadata for the category in question. For instance, the CLARIN.SI guidelines are listed under FTD because they provide quite a few instructions for this particular category; consider this snippet from the guidelines: "The resource description should be about half a page in length, and should describe the resource in terms of what it contains, approximate size, and basic structure of the data. Where relevant, it should give the envisioned use of the resource."

Instructions on language and resource name are included in roughly a third of the guidelines. The

[3] https://www.clarin.eu/content/assessment-procedure

[4] https://docs.google.com/spreadsheets/d/12vjFSLv6uoCcuj_jC9PS7V8_HjEOSmnJpiFfyRA_VJU/edit?usp=sharing

[5] See for instance the CLARIN.SI guidelines (https://www.clarin.si/repository/xmlui/page/deposit).

[6] See for instance the ARCHE guidelines (https://arche.acdh.oeaw.ac.at/browser/formats-filenames-and-metadata#metadata).

inclusion of instructions for documenting annotation fares the worst, as they are present only in 2 (12%) of the 17 English guidelines. For instructions on documenting more qualitative metadata, such as temporal and geographic coverage, the rate of inclusion in the guidelines is similarly infrequent. Overall, few repositories provide explicit instructions on documenting the structural metadata that is typically considered the most salient in resource documentation from the user perspective.

3 The Proposal for Guideline Refinement

Qualitatively, the existing instructions are generally lacking in detail and are quite uninformative for a user unfamiliar with CLARIN deposits. In the case of annotation, for instance, even the most detailed of the reviewed instructions are such that the depositor is simply prompted to specify whether the corpus is raw or annotated, but is not instructed to provide more fine-grained information on the key subcomponents of the annotation process itself, such as tagsets or annotation schemata.

We therefore propose a new set of depositing recommendations that will primarily focus on the qualitative aspects of the deposit, i.e. those that that are important for facilitating the (re)-use of a resource within the context of the wider SSH community. Concretely, we propose that repositories should adopt/adapt the following guidelines for documenting the four surveyed metadata categories as well as the free-text description:

Annotation. Unlike in the case of size, B-centres that use the DSpace system (Smith et al., 2003) for their repositories, i.e. LINDAT, ILC4CLARIN, CLARIN.SI, CLARIN:DK, CLARIN:PL, and SWE-CLARIN, do not provide the user with a special field for defining the annotation during the submission process. Depositors should therefore be explicitly instructed to provide a very brief summary of the annotation process, possibly as part of the free-text description. Depositors should ideally distinguish between linguistic (e.g., tokenisation, sentence segmentation, PoS-tagging, lemmatization, syntactic parsing, named entity recognition) and non-linguistic levels of annotation, which are often domain specific (e.g., gender annotation of speakers in parliamentary corpora). Optionally, information on additional subcomponents of the annotation process itself should be provided, such as the tagset used for morphosyntactic tagging, the class of named entities, and possible syntactic frameworks for syntactically parsed corpora (e.g., Universal Dependencies (De Marneffe et al., 2021) for dependency grammars), and the tools used to annotate the corpora.

Size. In line with the guidelines of the CLARIN:EL C-centre (The CLARIN:EL Technical Team, 2022), we propose that in case the corpus contains more than one modality (e.g., audio recordings and their written transcriptions in the case of spoken corpora), depositors should be explicitly instructed to provide size for each modality separately. Additionally, we propose that if the corpus is tokenised, depositors should make sure to provide both the word and token numbers – this distinction is not mentioned in any of the reviewed guidelines, but is crucial from the perspective of inter-resource comparability.

Resource Name. CLARIN.SI already requires that the "title [of a resource] should give a very short description of the resource, followed by its proper name, which is usually an acronym" and gives examples of resources following this naming convention, such as *Automatically sentiment annotated Slovenian news corpus AutoSentiNews 1.0* (Bučar, 2017). This is a crucial instruction because there are otherwise many deposited resources with non-transparent names, which is suboptimal from the perspective of an external user unfamiliar with the repository's catalogue.

Language. Depositors should be instructed to specify any possible and important characteristics of the resource's language(s) that are ambiguous in the language subcomponent of the metadata profile. For instance, if a bilingual corpus contains Slovenian and English texts, it should be specified which language corresponds to the original or translated texts or both, and relevant related characteristics such as the directionality of translations should be identified. If applicable (e.g., oral history corpora), the proportion of sources in the deposited corpora wrt. to their language should be clearly indicated.

Free-Text Description. The free-text description should mainly focus on the description of the resource itself rather than on background information such as funding; too much focus on describing the latter in lieu of the former is indeed a problem with many existing deposits, whereas the descriptions should be maximally informative to the research needs of an external user – see Lenardič and Fišer

(2020) for further discussion and examples. Depositors should be explicitly prompted to consider describing the following features of their resource or at least a subset thereof: *modality* (spoken, written, visual, etc.), time period (based on publication date), *geographic coverage*, *data sampling* (text types and their ratios; text sources and their ratios), *envisaged usage and relevant research domains*, and any important and unique *domain-specific characteristics* (e.g., participants' ages and L1s in learner corpora).

4 Conclusion

Depositing guidelines that instruct researchers to document their resources as comprehensively as possible are important in minimizing metadata gaps and inconsistencies at the stage before publication, which is important because post-hoc curation of published resources requires a significantly higher investment of effort and time.[7] For the next step, we plan to get in contact with repository administrators in order to discuss with them the possible adoption of the presented guidelines, which we believe can directly and tangibly increases the value of the individual deposits, individual repositories and the infrastructure as a whole.

Acknowledgments

The work described in this paper was funded by the Slovenian Research Agency research programme P6-0436: *Digital Humanities: resources, tools and methods* (2022–2027) and the DARIAH-SI and CLARIN ERIC (the Resource Families Project) research infrastructures.

References

Jože Bučar. 2017. Automatically sentiment annotated slovenian news corpus AutoSentiNews 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1109.

Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Garcia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Springer, Cham, Switzerland.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Koenraad De Smedt, Franciska de Jong, Bente Maegaard, Darja Fiser, and Dieter Van Uytvanck. 2018. Towards an open science infrastructure for the digital humanities: The case of CLARIN. In Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen, editors, *Digital Humanities in the Nordic Countries Conference (DHN)* 3, pages 139–151.

Marijn Koolen, Jasmijn Van Gorp, and Jacco Van Ossenbruggen. 2019. Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, 34(2):368–385.

Jakob Lenardič and Darja Fišer. 2020. The clarin resource and tools family. In Costanza Navaretta and Maria Eskevich, editors, *Proceedings of CLARIN Annual Conference 2020*.

John P. McCrae, Philipp Cimiano, Luca Matteis, Roberto Navigli, Victor Rodriguez Doncel, Daniel Vila-Suero, Jorge Garcia, Andrejs Abele, Gabriela Vulcu, and Paul Buitelaar. 2015. Reconciling heterogeneous descriptions of language resources. In *Proceedings of the 4th Workshop on Linked Data in Linguistics*, pages 39–48, Beijing. Association for Computational Linguistics.

Jan Odijk. 2019. Discovering software resources in CLARIN. In *Selected Papers from the CLARIN Conference, Pisa, 8–10 October, 2018*, Linköping Electronic Conference Proceedings 159, pages 121–132, Linköpings universitet. Linköping University Electronic Press.

MacKenzie Smith, Mary Barton, Mick Bass, Margret Branschofsky, Greg McClellan, Dave Stuve, Robert Tansley, and Julie Harford Walker. 2003. DSpace: An open source dynamic digital repository. *D-Lib Magazine*, 9(1). http://doi.org/10.1045/january2003-smith.

The CLARIN:EL Technical Team. 2022. Clarin release 1. https://clarin-platform-documentation.readthedocs.io/_/downloads/en/stable/pdf/, accessed 12 April 2022.

[7] Note, however, that we do not suggest that the guidelines presented in this paper should become obligatory CLARIN requirements for new deposits. The proposed guidelines should rather be thought of as recommendations that guide repository administrators when assessing submissions from the perspective of (re)use in research.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3. `http://doi.org/10.1038/sdata.2016.18`.

Peter Wittenburg, Dieter Van Uytvanck, Thomas Zastrow, Pavel Straňák, Daan Broeder, Florian Schiel, Volker Boehlke, Uwe Reichel, and Lene Offersgaard. 2019. Checklist for clarin b centres. `http://hdl.handle.net/11372/DOC-78`.

The Resource Publishing Pipeline of the Language Bank of Finland

Ute Dieckmann
ute.dieckmann@helsinki.fi

Mietta Lennes
mietta.lennes@helsinki.fi

Jussi Piitulainen
jussi.piitulainen@helsinki.fi

Jyrki Niemi
jyrki.niemi@helsinki.fi

Erik Axelson
erik.axelson@helsinki.fi

Tommi Jauhiainen
tommi.jauhiainen@helsinki.fi

Krister Lindén
krister.linden@helsinki.fi

Department of Digital Humanities
University of Helsinki, Finland

Abstract

We present the process of publishing resources in Kielipankki, the Language Bank of Finland. Our pipeline includes all the steps that are needed to publish a resource: from finding and receiving the original data until making the data available via different platforms, e.g., the Korp concordance tool or the download service. Our goal is to standardize the publishing process by creating an ordered check list of tasks with the corresponding documentation and by developing conversion scripts and processing tools that can be shared and applied on different resources.

1 Introduction

The Language Bank of Finland (Kielipankki, "The Language Bank") is the collection of services coordinated by FIN-CLARIN, the Finnish consortium of universities and research organizations. Various types of resources can be deposited in the Language Bank, including for instance text and speech corpora, lexicons and terminologies, and many kinds of data sets produced by research projects.

The Language Bank supports public, academic as well as restricted license categories and offers multiple services for providing access to different resource variants. Since the publication framework is complex and not yet sufficiently automatic, depositors cannot directly upload their resources. We need to support them in clearing the licenses and in converting, annotating and describing their data.

Currently, more than 250 resources are available via the Language Bank of Finland. About 100 resources are listed as forthcoming, and more are added every month. Before implementing the publishing pipeline, every team member involved in the process of publishing resources had their own workflows and conversion scripts. This tended to result in slight inconsistencies in the published data. Monitoring the current state of a given resource within the publication process was not always easy. Some tasks, like parsing the data for publication in Korp, were carried out by only one person in the team, making processes very dependent on this person's availability and time.

Ideally, all resources published via CLARIN services should meet the FAIR standards (findable, accessible, interoperable, reusable)[1]. By creating a shared and well-documented workflow and by using common tools, we aim to ensure that all resources and their future versions are processed, published and maintained in a consistent, transparent and interoperable way.

2 Means of Publication

Resources published in the Language Bank of Finland may be available via the online concordance tool Korp, developed by Språkbanken, the Swedish Language Bank (see Borin, Forsberg & Roxendal, 2012),

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

and adapted for the Language Bank of Finland[2]. Resources can also be downloaded via the download service[3] of the Language Bank. Several variants can be offered of the same resource. For the users' convenience, copies of selected versions of the downloadable corpora are also accessible in the computing environment at CSC – IT Center for Science[4]. Lexical resources can be made available via Sanat[5]. For storing the internal backup copies of each resource, we use IDA, the research data storage service organized by the Ministry of Education and Culture in Finland[6].

The resource group page of a given resource on the website (the "portal") of the Language Bank lists all available versions of the resource, including links to their metadata records, their access locations, and further information. A link to the resource group page can be found in the metadata record of each version of the resource. Following the example of CLARIN Resource Families[7], we also offer a portal page where the resource groups in the Language Bank are categorized under CLARIN-style families.

2.1 Access Rights

The Language Bank aims to provide resources as openly as possible. Many resources can be made publicly available (CLARIN PUB license category). However, access restrictions are often necessary for protecting copyrighted content or personal data. Some resources are licensed for academic use only (ACA), and they may be accessed by signing in with credentials issued by the user's home institution. Furthermore, the Language Bank is able to distribute resources under restricted licenses (RES), in which case users can apply for individual access rights in the Language Bank Rights (LBR) service[8]. LBR currently supports federated login and user identities via CLARIN[9] or Eduuni[10].

Unless the original material has been previously available under a public license, the licenses of individual resources in the Language Bank are based on agreements with the right holders and, in the case of resources that contain personal data, with the data controllers.

In some cases, it is possible to offer several variants of the same resource under different licenses. For instance, since speakers might be identifiable based on their voice, audio speech recordings often need to be protected, e.g., by restricting access to them. However, anonymized or pseudonymized transcripts could be available under a less restricted license for purposes where audio is not needed.

3 Tasks within the Publishing Pipeline

The process of publishing an individual corpus usually involves 3–4 people in the Language Bank. In case of an exceptionally simple and well described dataset with no licensing issues, it does not take more than one or two working days to publish the source data for download. If intense license discussions and several different means of publication are required, the process can take up to 60 working days.

For each new resource, we maintain a check list[11] of the tasks in the shared pipeline that are relevant for the resource in question. The list is used for keeping track of the status of the resource during the publishing process. Some tasks on the list are mandatory, whereas some are applicable to specific resources only. According to the type of task, which can be for example administrative or technical, work can be assigned to the person with the required skills.

3.1 Entering a New Resource to the Pipeline of the Language Bank of Finland

When a researcher or a research group creates a new resource that they wish to make available to other researchers or publicly, they are first asked to submit the most important details regarding the resource by filling in an e-form[12]. The Language Bank then creates a preliminary metadata record on the local

[2] https://korp.csc.fi
[3] https://www.kielipankki.fi/download
[4] https://www.csc.fi
[5] https://sanat.csc.fi
[6] https://ida.fairdata.fi
[7] https://www.clarin.eu/resource-families
[8] https://lbr.csc.fi
[9] https://www.clarin.eu/content/clarin-identity-provider
[10] https://info.eduuni.fi/en/services/eduuni-id
[11] https://github.com/CSCfi/Kielipankki-utilities/blob/master/docs/corpus_publishing_tasklist.md
[12] http://urn.fi/urn:nbn:fi:lb-2021121422

META-SHARE repository[13]. The preliminary metadata are checked together with the depositor. The details can be updated and amended later. The metadata are automatically harvested from META-SHARE by other services, e.g., the Virtual Language Observatory[14] maintained by CLARIN.

For publications, researchers may need a persistent reference to their resource before it is made available by the Language Bank. Since unofficial links should be avoided in citations, the Language Bank assigns a persistent identifier (PID) to the metadata record as soon as the resource exists and has been sufficiently well described. This PID is the citable and primary identifier for the resource (for details on how the Language Bank uses PIDs, see Matthiesen & Dieckmann, 2019). At this point, the resource is also added to the list of forthcoming resources on the website of the Language Bank.

3.2 Clearing the License for the Resource and Acquiring the Source Data

Unless the resource has previously been published under an open license, the Language Bank and the depositor negotiate on the license for distributing the resource. If the resource contains copyrighted content, additional steps may be needed to obtain permissions from the copyright holders. If the resource includes personal data, the data controller is involved in the deposition agreement. In this case, the end-user license will include the condition +PRIV, and all users who access the resource via the Language Bank will be required to comply with the resource-specific data protection terms and conditions.

The Language Bank uses a generic deposition license agreement template[15]. In order to discuss the details, a meeting with the depositor is often needed. When an agreement is reached, the end-user license is published in the portal. Using PIDs, the metadata record will refer to the license page and vice versa.

After receiving the source data from the resource depositor, they are checked for format and validity, and a description of the contents is added for internal use. A backup copy of the data is stored in IDA.

3.3 Publishing the Source Data in Download

Since the data conversion process tends to take time, the very first version of a corpus to be published is usually the source data that is made available for download. In this version, the original content is not modified. However, the metadata and license information must be available and up to date.

A PID is added to the metadata record, and a resource group page, which also gets a PID, is created and linked with the corresponding metadata. The source version of the resource, and possibly other planned versions of the resource, are added to the list of forthcoming resources in the portal, to keep the corpus owners and possibly interested researchers informed. A PID for the download location is requested.

To prepare the resource for download, the source data is packaged into one or more zip files as agreed with the corpus depositor. In case the license of the resource is RES, a record is created on the LBR system in order to control access to the download location. Similarly, if the license is ACA, academic user login will be required to download the resource. After testing the zipped packages, they are uploaded to the download service, together with a README text file containing basic information on the resource and a LICENSE text file offering information on access rights for this resource. The metadata record and the resource group page can then be updated, adding the access location PID.

To finalize the publishing of a resource, it is added to the list of published resources in the portal. A news item is published in the portal to inform interested researchers about the new resource. The depositor is informed about the publication as well. The download package is uploaded to IDA, and in selected cases the unpacked source data is also made available in CSC's computing environment.

3.4 Publishing the Data in Korp

If a resource is meant to be made available in Korp, the data is tokenized, parsed (if a parser is available for the language in question), and possibly extended with additional annotations (name annotations, sentiment annotations, identified languages). The first steps of publishing a resource are similar, however, regardless of the means of publication. A metadata record for the Korp version is created or updated and PIDs are assigned to the metadata record and to the access location. In case the license of the resource is RES, an LBR record is created.

[13] https://metashare.csc.fi

[14] https://vlo.clarin.eu

[15] https://www.kielipankki.fi/support/dela/

The format of the original data may differ between corpora. It can be for example plain text, PDF, RTF, a tabular format such as CoNLL-X, or an XML format such as TEI. The first aim is to convert this data to a simple form of XML, which must be UTF-8 encoded Unicode. This task is carried out with individually developed scripts and usually is the most time-consuming, depending on the format of the original data. The basic idea is to segment the content of the original files so that plain text is inside text and paragraph tags, which can include descriptive attributes. These files with a relatively simple structure are then used as input for further processing tools.

The next step is the tokenizing process where the paragraphs are segmented into sentence elements and tokens. The output format of the tokenizer is VRT (VeRticalized Text), the input format for the IMS Open Corpus Workbench (CWB) software underlying Korp. In addition, we have extended the VRT format with a comment that provides names for the otherwise positional attributes of tokens.

It is possible for the Language Bank to apply further tools on the VRT data to add any desired annotations, while preserving the sentence and token boundaries and previous annotations. For instance, information about the languages used in the text could be added by running a language identifier such as HeLI-OTS (Jauhiainen & Jauhiainen, 2022) that includes language models for 200 languages.

For Finnish and other languages with a parser and named-entity recognizer available, the parsing process is carried out on the validated VRT data. For years, we have been using an early version of the Turku dependency parser for Finnish, developed by the Turku NLP group and adapted for VRT. We are currently adopting their new neural parser[16] along with the Universal Dependencies annotation model.

One single but complex script handles the processing of VRT files to create a Korp corpus package containing CWB data files and Korp MySQL database import files. The resulting package is then installed on the Korp server and a corpus configuration is added. When the test instance meets expectations, the corpus is published in Korp as a beta test version. The new corpus is announced in the Korp news desk as well as in the portal. The beta status is removed after two weeks unless requests for changes appear during this period. Finally, a copy of the Korp corpus package is stored in IDA.

After publishing a resource in Korp, the VRT data is usually extracted from Korp and published in the download service in order to provide similar versions of the data via both channels. The VRT version of a resource is published in the download service in the same way as the source data.

4 Conclusions

Currently, the Language Bank of Finland provides researchers with access to over 250 resources, and many more are forthcoming. The licensing and publishing process of each resource takes time and effort and tends to require various kinds of expertise. Based on our experience, we have identified a number of tasks that are relevant when publishing most types of resources, resulting in a check list and modular documentation[17] offering instructions for the individual tasks. Although this pipeline is still under development, the general workflow has already proven useful for managing and monitoring the publication process more efficiently. We aim to automatize and document the process even further to enable resource depositors to take a more active role in preparing their data. We believe that by comparing and sharing good practices with other CLARIN centres, it is possible to support researchers even better.

References

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

Tommi Jauhiainen and Heidi Jauhiainen. (2022). HeLI-OTS 1.3 (1.3). Zenodo. https://doi.org/10.5281/zenodo.6077089

Martin Matthiesen and Ute Dieckmann. (2019). A PID is a Promise – Versioning with Persistent Identifiers. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 103–112.

[16] http://turkunlp.org/Turku-neural-parser-pipeline/

[17] https://github.com/CSCfi/Kielipankki-utilities/tree/master/docs

TEI and Git in ParlaMint: Collaborative Development of Language Resources

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

Matyáš Kopp
Faculty of Mathematics and Physics
Charles University
Prague, Czech Republic
kopp@ufal.mff.cuni.cz

Abstract

This paper discusses the encoding, validation and development of language resources of the completed ParlaMint I and on-going ParlaMint II CLARIN projects, which centre on the collaborative development of TEI-encoded corpora of parliamentary proceedings. We introduce the use of TEI ODD to write the encoding guidelines and formal XML schemas for validation. We motivate and explain using Git to develop of encoding schemas and language resources. Apart from revision control, issues and publishing documentation, we also emphasise GitHub actions with their ability to integrate program code execution into the data submission process. The paper is written with a view to introducing SSH scientists to the two environments, as they can be valuable items in the toolbox for compiling language resources, especially in a collaborative setting.

1 Introduction

ParlaMint is a CLARIN ERIC supported project which, among other tasks, aims to produce a set of comparable and richly annotated corpora, involving a joint effort of a large number of partners. The concluded ParlaMint I project (2020–2021) already developed corpora containing transcriptions of the sessions of 17 European national parliaments in the time-span 2015–2021 (Erjavec et al., 2022). These corpora are about half a billion words in size, contain rich metadata about the 11 thousand speakers, and are linguistically annotated. The on-going ParlaMint II project (2022–2023) plans to extend existing corpora with newer data and add corpora for 12 new, also regional, European parliaments. It will also enhance the corpora by providing machine translations to English, and, for a selected subset of corpora, add speech data, as well as work on the wider use of the corpora.

For these reasons, it is important to have a robust but easily maintainable encoding, along with documentation, as well as automated validation and conversion procedures for the project's corpora. Indeed, one of the first tasks in ParlaMint II is to re-evaluate and extend these aspects of the project, and the two main environments used are TEI and Git. We discuss these two aspects of the projects, both in the light of future improvements, and also to introduce them to DH scholars for their own use.

2 Using TEI

The encoding of the ParlaMint corpora follows the Parla-CLARIN recommendations for encoding parliamentary corpora (Erjavec and Pančur, in print), which is itself a customisation of the TEI Guidelines (Consortium, 2022)[1]. A TEI customisation is specified in a TEI ODD document, which serves a double function: it contains the prose guidelines, and the formal schema of the customisation, using the TEI ODD schema specification language. With the TEI XSLT

[1] https://tei-c.org/guidelines/P5/

stylesheets the prose guidelines can be converted to HTML for reading, while the ODD schema specification can be converted into one of the standard XML schema languages, such as the ISO standard RelaxNG, and such an XML schema is then used for formal validation of the corpora. The design of the Parla-CLARIN recommendation was inspired by previous similar efforts, in particular the TEI Lex-0 encoding recommendations for dictionaries (Tasovac et al., 2018)[2] and the TEI schema for the multilingual ELTeC corpus containing 100 historical novels for a number of languages (Burnard et al., 2021; Schöch et al., 2021)[3].

Although the ParlaMint corpora conform to the Parla-CLARIN schema, we required, in order to ensure interoperability, a much more constrained encoding than the quite general one of Parla-CLARIN. To this end, we developed in ParlaMint I a bespoke RelaxNG schema directly, without using the ODD mechanism. The advantage of this approach is that the schema expresses exactly the kinds of constraints that we wished to make, while the disadvantage is that there were no guidelines accompanying the schema and, because the schema was not derived from a TEI ODD, there is no formal guarantee that the schema is still Parla-CLARIN compatible, or, indeed, compatible with TEI[4].

For these reasons, we are in ParlaMint II, developing a ParlaMint ODD. So far, we have written the prose guidelines for ParlaMint corpora, and partially developed the ODD schema. We are also working on the documentation of individual elements and attributes in the ODD schema, i.e. changing the default glosses and examples of use of the elements as they appear in the TEI to ParlaMint specific ones.

It should be noted that the ParlaMint schema (either the bespoke or the ODD-derived RelaxNG) is only the first step in the validation of the ParlaMint corpora. We have also developed an XSLT script that performs validation regarding the textual content of some elements, and checks that redundant meta-data (i.e. meta-data which is encoded more than once in the corpora but which makes it easier to process the corpora further) is not contradictory. Furthermore, as the corpora are converted to other formats, such conversions can also expose various errors in the corpus encoding. Finally, as the corpora are mounted on concordancers, the corpus compilation process, or the actual analysis of corpora on the concordancers can also reveal bugs in the data; all these quality control mechanisms have been used in ParlaMint I and are planned to be used in ParlaMint II.

3 Using Git

Git has become the revision control system of choice for most software development projects, and has also proved its worth in the (collaborative) development of language resources, with the most prominent example being the Universal Dependencies treebanks and annotation guidelines (de Marneffe et al., 2021)[5]. It has also been used for the development of TEI customisations, e.g. the already mentioned TEI Lex-0 and ELTeC, for the latter used not only of the schema, but of the corpora as well.

Apart from support for collaborative development with transparent versioning and attribution and simple comparisons of files, Git hosting platforms, such as GitHub, support social media aspects of development, in particular posting and discussing issues, commenting on commits or pull requests, and a Wiki space. It is also possible to directly publish the documentation of a project using GitHub Pages. Finally, running scripts at a particular point in the Git workflow is supported by GitHub Actions. All these features lead to a more controlled and better documented development process.

We had already used Git for the development and publishing of the Parla-CLARIN schema,

[2] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[3] https://github.com/COST-ELTeC

[4] In order to ensure such compatibility, each ParlaMint corpus is validated both against the bespoke ParlaMint RelaxNG schema, as well as against the Parla-CLARIN schema.

[5] https://universaldependencies.org/

where the TEI ODD is maintained on GitHub[6], the guidelines are published as GitHub pages[7], and technical instructions for using or further developing the schema are available on the GitHub Wiki[8]. In ParlaMint I, as well, the project development was to a large extent done on GitHub[9]. The Git contained the latest RelaxNG schemas for the corpora and the complete validation or transformations scripts, written mostly in XSLT (and some Perl). Problems with the proposed encoding schema were often discussed through GitHub issues, while problems with individual corpora were communicated mostly by email.

While Git(Hub) is well suited for developing, storing and publishing software tools, schemas, and (by definition rather small) hand-annotated corpora, the complete ParlaMint corpora are, in practice, too large to be stored in Git, and, even more so, GitHub. The reason is the sheer size of the corpora (over 240GB for ParlaMint I) and number of files (almost 160,000 files), as well as the fact that, say, a new round of automatic annotation changes almost all lines in most files, making such a commit a very slow process. Therefore we here opted for a compromise, namely, we developed a script to extract only small samples from complete individual corpora, and maintain only these samples, also in derived formats, on GitHub. These samples can be directly viewed on GitHub, and this give an impression of how the corpora are structured.

In ParlaMint II, the first step was to update the Parla-CLARIN GitHub to reflect the ParlaMint best practice, while the ParlaMint GitHub was extended with pages[10] publishing the ParlaMint encoding guidelines. As ParlaMint II has now (at least) 30 partners, we hope to identify and resolve all problems through GitHub issues, rather than via email, so that problems are documented, can be discussed and the solution linked to a commit.

As we cooperate with many partners that are supposed to add their sample data with pull requests, a validation procedure for newly inserted data using GitHub Actions[11] has been developed. Furthermore, when the pull request with valid data is merged into the correct branch, the TEI files are sampled, and derived formats are added to the repository. This approach has several benefits: there is no need for partners to carefully sample the data themselves, they do not need to compile derived formats, and we can be sure that the derived files are always up to date with corresponding TEI files.

Of course, it will also be necessary to validate and convert the complete corpora. For local processing we have developed a validation procedure that uses the Unix `make`[12] tool. The Makefile is self-documenting for easier use, i.e. running `make` without arguments prints a list of the available targets, which e.g. check installed prerequisites, validate the corpus against the ParlaMint and Parla-CLARIN schemas, perform advanced content validation, and convert a corpus to derived formats.

4 Conclusions

We attempted to show how TEI can be used to specify the encoding of language corpora (or other types of language resources), providing both the guidelines of those wishing to encode the corpora, as well as XML schemas that are used to formally validate their encoding.

We also presented the Git environment which is well suited for controlled and distributed development and publishing of not only the guidelines and schemas, but also the language resources themselves. As mentioned, the size of the produced ParlaMint corpora precludes storing them in their entirety on Git but this is not the case for smaller language resources, particularly manually annotated ones. Fully mastering Git is also not a simple process, esp. for the typical researcher of the target SSH community. However, with an appropriate set-up, such as we have

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

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

[8] https://github.com/clarin-eric/parla-clarin/wiki

[9] https://github.com/clarin-eric/ParlaMint

[10] https://clarin-eric.github.io/ParlaMint/

[11] https://docs.github.com/en/actions

[12] https://www.gnu.org/software/make/

attempted to provide for ParlaMint, we believe that with only basic knowledge partners can successfully submit their data sample and check if they validate, while the validation of the complete corpus still rests on the editors.

In our future work, we plan to continue working on the ParlaMint ODD, which will also be updated as the new corpora, with new problems, appear in ParlaMint II. New types of annotations (e.g. semantic annotation) and types of resources (machine translated corpora, speech data) will also need to be supported, so we will extend the ODD to support and document them. We will also continue working on the Git(Hub) environment, including a tutorial for ParlaMint partners on its use.

The development of the ParlaMint corpora is also currently still rather centralised. In the longer perspective, we would like to encourage anyone that would wish to produce a ParlaMint-compatible corpus to be able to do so independently, for which we have to make the set-up (even) more flexible.

We believe that both TEI and especially Git – and the possibilities of combining the two – are not as well known in the SSH community as they should be, and that learning about them and adopting them into the work process could go a long way in making the development of encoding guidelines and language resources a much smoother and more controlled process, also leading to better reproducibility, a point that is very relevant to the goals of the CLARIN infrastructure.

Acknowledgements The authors would like to thank the three anonymous reviewers for their helpful comments and suggestions, and all the partners in ParlaMint I and II for their work on the corpora and suggestions for improving the ParlaMint schema and tools. The work presented in this paper was supported by the CLARIN ERIC projects ParlaFormat (2019) and "ParlaMint: Towards Comparable Parliamentary Corpora" (2020–2021 and 2022–2023), by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ, and individual project partner grants.

References

Lou Burnard, Christof Schöch, and Carolin Odebrecht. 2021. In Search of Comity: TEI for Distant Reading. *Journal of the Text Encoding Initiative*, (14).

TEI Consortium. 2022. *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Tomaž Erjavec and Andrej Pančur. in print. The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings. *Journal of the Text Encoding Initiative.*

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation.*

Christof Schöch, Roxana Patraș, Diana Santos, and Tomaž Erjavec. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, (1). http://doi.org/10.3828/mlo.v0i0.364.

Toma Tasovac, Laurent Romary, Piotr Banski, Jack Bowers, Jesse de Does, Katrien Depuydt, Tomaž Erjavec, Alexander Geyken, Axel Herold, Vera Hildenbrandt, Mohamed Khemakhem, Boris Lehečka, Snežana Petrović, Ana Salgado, and Andreas Witt. 2018. TEI Lex-0: A baseline encoding for lexicographic data. Version 0.9.1. Technical report, DARIAH Working Group on Lexical Resources.

Analysing Changes in Official Use of the Design Concept Using SweCLARIN Resources

Lars Ahrenberg, Daniel Holmer, Stefan Holmlid, Arne Jönsson
Department of Computer and Information Science
Linköping University, Linköping, Sweden
firstname.lastname@liu.se

Abstract

We show how the tools and language resources developed within the SweClarin infrastructure can be used to investigate changes in the use and understanding of the Swedish related words *arkitektur*, *design*, *form*, and *formgivning*. Specifically, we compare their use in two governmental public reports on design, one from 1999 and the other from 2015. We test the hypothesis that their meaning has developed in a way that blurs distinctions that may be important to stakeholders in the respective fields.

1 Introduction

What is the relation between architecture and design? As concepts in the minds of speakers or as professions where stakeholders sometimes compete and sometimes join forces to achieve their goals? In this paper we try to answer such questions using the resources developed for the analysis of Swedish by Språkbanken Text and distributed through the SweClarin portal. More specifically we want to study whether there is a change in the denotations and connotations of four related words: *arkitektur*, 'architecture', *form*, 'form', *formgivning* (cf. German *Formgebung*) and *design*. In particular, are there changes in their use and, perhaps, signs of a convergence? Its rationale is a hypotheses from colleagues working in design that there has been an increased effort to place architecture and design under the same umbrella, not least from the side of the Swedish government, and that this development has been detrimental for the design field.

Dictionary definitions of the four words vary. According to one of them[1], the word *design* was first observed in Swedish in 1948. It is defined there as *konstnärlig formgivning*, 'artistic form giving' using the older term *formgivning*. Over the years *design* has established itself as a synonym of *formgivning*, and also, as will be shown, become the more frequently used of the two. The word *arkitektur*, 'architecture', on the other hand is defined as a science with a related concrete meaning as 'artistic and technical design of buildings'. Thus, *arkitektur* can be defined in terms of *design* but also in other terms. Nowadays, also *design* can be studied at universities as a separate subject. The word *form* has many meanings, one of them being 'artistic form'. It is used in this sense by the private organisation *Svensk Form*, 'Swedish Form', established in 1845, and its journal, simply named *Form*.

The studies presented in this paper make use of the Sparv text analysis pipeline[2] (Borin et al., 2016), the senSALDO[3] sentiment lexicon (Rouces et al., 2019), and the Swedish Culturomics Gigaword Corpus (Rødven Eide et al., 2016).

2 Historical background

In 1997 the Swedish government proposed an action program for an area identified as *arkitektur och formgivning*. Two years later an official governmental report (SOU), entitled *Mötesplats för form och*

[1] Nationalencyklopedins Ordbok, 1995 edition.

[2] https://spraakbanken.gu.se/sparv/#/sparv-pipeline

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

design, ’A meeting-place for form and design’[4] proposed a new initiative for design. The report argued, however, that it was reasonable that its proposals had a clear connection to architecture, as its proposals related to buildings and building sites and an upcoming ’Year of Architecture’, referring to the year 2001.

A few years later, in 2009, the Swedish Museum of Architecture was given a new responsibility to cover also ”other fields of design” and its name was changed to ArkDes: The Swedish Centre for Architecture and Design. Its mission is ”to increase knowledge of and cultivate debate about how architecture and design affect our lives as citizens.”[5] More recently a new SOU-report was requested which was ready in 2015. With the title *Gestaltad livsmiljö: en ny politik för arkitektur, form och design*, ’Shaped habitat: a new policy for architecture, form and design’, it brought the three concepts architecture, form, and design closer together and proposed the establishment for a new public body dealing with them jointly.

3 Data

In addition to the two SOU reports mentioned above we have used the news sections of the Swedish Culturomics Gigaword Corpus (Rødven Eide et al., 2016) from relevant time periods for comparisons. Frequencies for the terms of interest in the different datasets are shown in Table 1 and Figure 1.

	arkitektur	design	formgivning	form	arkitektur, form och design	all tokens
Gw 1990-99 (News)	793	1,008	204	10,421	0	60,037,845
Gw 2010-2015 (News)	1,595	4,042	303	28,102	0	168,998,305
SOU 1999:123	43	334	139	180	0	37,880
SOU 2015:88	318	317	26	301	0	47,345
- retokenized*	148	163	26	131	170	46,665

*The triad *arkitektur, form och design* considered as one unit, see Section 4

Table 1: Frequencies of the investigated words in different corpora.

We can see here that the ratio of *design* to *formgivning* is changing rather rapidly. For the 1990s the ratio is about 5:1, rising to to 13:1 for the period 2010-2015. Also in the SOUs the word *formgivning* is used less. Not only is it losing ground to *design* but also to *form*.

The word *arkitektur* is infrequent in the SOU from 1999 as the main topic of that report is design. However, there are many indirect references to architecture which we can see if we take into account compounds and derivations. In particular there are plenty of references to *Arkitekturmuséet*, ’The Museum of Architecture’, and *Arkitekturåret*, ’The Year of Architecture’.

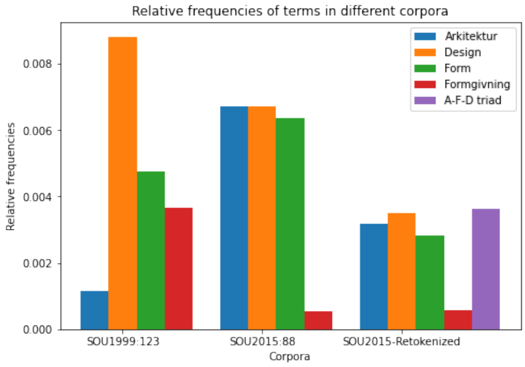


Figure 1: Bar chart showing relative frequencies in three corpora.

4 Analyses

Word distribution. To understand how the terms are used in general language, we looked at their distribution in the news sections of the Swedish Gigaword Corpus (Rødven Eide et al., 2016), for the periods 1990-99 and 2010-2015. Using word embeddings derived with the Gensim Word2Vec framework, we can observe the following:

- *design* and *formgivning* are close (synonyms) for both periods. The word *grafisk*, ’graphical’, a common attribute to both terms, is about equally close.
- In the 1990:ies *formgivning* is a close neighbour to *arkitektur*, while *design* is further away. In the period 2010-15 the situation is reversed. In this period, *design* and *konst*, ’art’ are competing for the place as closest neighbour to *arkitektur*.

[4] SOU 1999:123

[5] https://arkdes.se/en/about-us/

- The word *form* does not turn up in the close vicinity of any of the other words. This is due to its many other, more common meanings such as ’type’, ’sort’, ’shape’, ’state’, ’mould’.

In order to compare the semantic space of the studied terms in the two reports, we used the temporal word analogies method as suggested by (Szymanski, 2017). This method works by transforming two vector space models to a common vector space, which acts like a link between the models, enabling the comparison of word vectors between two otherwise independent models. Thus, we can investigate shifts between the models, in the form of *”which word X in model M1 correspond to word Y in model M2?”*.

We trained a Word2Vec-model for each of the reports and applied the temporal word analogies technique to search for differences in usage of the studied terms, *architecture, design*, and *form*. Although the models themselves showed some differences when extracting and manually inspecting their most similar words, this method did not reveal any semantic shift of any of the studied terms between the two reports. It is possible, however, that this is due to the relative small size of the data and vocabularies used.

Part of the reason why the words turn up as close neighbours in vector space is that they are often coordinated, in pairs, triads or even longer ones that include words such as *konst*, ’art’, and *hantverk*, ’crafts’. We also see trends of concept building via these coordinations. First in the name of the ArkDes Museum, and later, in the SOU from 2015, where the triad *arkitektur, form- och design* is very frequent and treated as such in proposals as well as in reactions to the proposals. In fact, out of the instances found in SOU 2015:88 as many as 170, or more than 50% for all three words, appear as part of this triad. For this reason we created a version of the report where this triad was treated as a single token (see Table 1).

Topic modelling of the SOUs. We have applied topic modelling to the reports to see whether they differ in their distribution of topics, using the Gensim package (Řehůřek and Sojka, 2010) on parsed versions of the reports. The number of topics per model were chosen to maximise the coherence score C_v (Röder et al., 2015), which resulted in the model for the 1999 SOU having 16 topics, and the 2015 SOU having 14 topics.

We could see for the 2015 SOU, that for the topics where *design* is among the 10 most relevant terms, so is *arkitektur*, and vice versa. For the majority of topics where this happens, *form* is also among the 10 most relevant terms.

Sentiment analysis of the SOUs. For sentiment analysis we used the Swedish SenSALDO 0.2 sentiment lexicon (Rouces et al., 2019) with sentiment scores -1, 0 and +1, and Vader (Hutto and Gilbert, 2014). What makes SenSALDO 0.2 unique in a Swedish context is that it assigns different sentiment values to different senses of a word, for instance the Swedish word *fara* can mean *danger* or *go (away)* where the former has a negative sentiment and the latter is neutral. SenSALDO comprises 12287 lexical entries where 8893 are unique words. Word sense disambiguation with the SenSALDO 0.2 lexicon is provided by the Sparv pipeline. Vader produces a compound score for each sentence, by summing the valence scores of the words according to their identified sense, and normalize this sum to be between -1 and +1.

The mean sentiment for the 1999 SOU is 0.106 and for 2015 it is 0.155. The difference is significant, $p < 0.001$ (all tests use Welch’s t-test). Thus, the 2015 SOU uses overall a more positive tone.

	Form		**Formgivning**		**Architecture**		**Design**	
	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment	Sentences	Sentiment
1999	150	**0.1231***	127	0.1098	42	0.173	323	**0.1256***
2015	238	**0.2019***	21	0.2218	246	0.2236	277	**0.2003***

Table 2: Concept based sentiment. Number of sentences and mean concept sentence sentiment. *Significant, $p < 0.001$

We have also investigated the sentiment for each of the concepts in focus *architecture, design, form*, and *formgivning*. For each concept, sentiment is computed if the concept occurs in the sentence. The result is presented in Table 2 showing that the 2015 SOU has a more positive attitude towards the concepts *form* and *design*.

However, if we consider the triad *arkitektur, form- och design* and filter out all sentences containing it, none of the sentiments differ significantly, i.e. the significant difference for *form* and *design* in one way or another depends on the triad. The triad is not used much in 1999, only 3 occurrences, so we also compared the triad to the other concepts for 2015 and then it turns out that the difference in sentiment for *form*, 0.1490, is significantly different from the triad, 0.2745, $p < 0.001$, and for *design*, 0.1678, $p < 0.001$. That is, the triad is presented with a more positive sentiment in the 2015 SOU.

5 Conclusions

We have compared two public government reports concerning the design concept and its status as a domain for public support, financially and structurally. The first one was published in 1999 and the second one in 2015. In particular, we look at relations between the concepts of design and architecture, as they are presented in the reports and in comparable news corpora. We can see indications of a semantic convergence of the word *design* with the word *arkitektur*, especially in the latter report. Also, in the reports as in the news corpora it seems to overtake the role of the older word *formgivning*. Moreover we see that the compound term *arkitektur, form and design*, that is so frequent in the 2015 report, is significantly more positively described there, compared to *form* and *design* as individual terms, seemingly underlining a supporting policy that takes an integrative approach.

We also conclude that the tools and language resources available in the SweClarin infrastructure for analysis of Swedish texts enable comparisons of language use also over such short time spans as 20 years. In particular, we exploited the ability of the Sparv parser to identify word senses for sentiment analysis, and the Culturomics Gigaword Corpus for comparing official government reports with general language. Sentiment analysis seems to be the method that provides the most reliable results in our case, while the results from topic modelling and temporal word analogies are more uncertain due to the small dataset.

References

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016.*

C.J. Hutto and E.E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI.*

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

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.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2019. Sensaldo: Creating a sentiment lexicon for Swedish. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 4192–4198.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The swedish culturomics gigaword corpus: A one billion word swedish reference dataset for nlp. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, Krakow, Poland.*

Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 448–453, Vancouver, Canada, July. Association for Computational Linguistics.

A Snapshot of Climate Change Arguments: Searching for Recurring Themes in Tweets on Climate Change

Maria Skeppstedt[1,2], **Robin Schaefer**[3]
[1]Institute for Language and Folklore, Stockholm, Sweden
[2]Centre for Digital Humanities Uppsala, Department of ALM, Uppsala University, Sweden
maria.skeppstedt@abm.uu.se
[3]Applied Computational Linguistics, University of Potsdam, Potsdam, Germany
robin.schaefer@uni-potsdam.de

Abstract

We applied the topic modelling tool Topics2Themes to a collection of German tweets on the subject of climate change, the GerCCT corpus. Topics2Themes is currently being further developed and evaluated within Språkbanken Sam, which is a part of SWE-CLARIN. The tool automatically extracted 15 topics from the tweet collection. We used the graphical user interface of Topics2Themes to manually search for recurring themes among the eight tweets most closely associated with the topics extracted. Although the content of the tweets associated with a topic was often diverse, we were still able to identify recurring themes. More specifically, 14 themes that occurred at least three times were identified in the texts analysed.

1 Introduction

The Intergovernmental Panel on Climate Change (IPCC) concludes that:

> "It is unequivocal that human influence has warmed the atmosphere, ocean and land. Widespread and rapid changes in the atmosphere, ocean, cryosphere and biosphere have occurred. [...] Global surface temperature will continue to increase until at least mid-century under all emissions scenarios considered. Global warming of 1.5°C and 2°C will be exceeded during the 21st century unless deep reductions in CO2 and other greenhouse gas emissions occur in the coming decades." (IPCC, 2021)

Despite that, there are still many debates on climate change that do not mainly evolve around the subject of climate change mitigation. So, what subjects are then for instance debated instead? What themes do frequently occur in claims and arguments presented in debates about the climate?

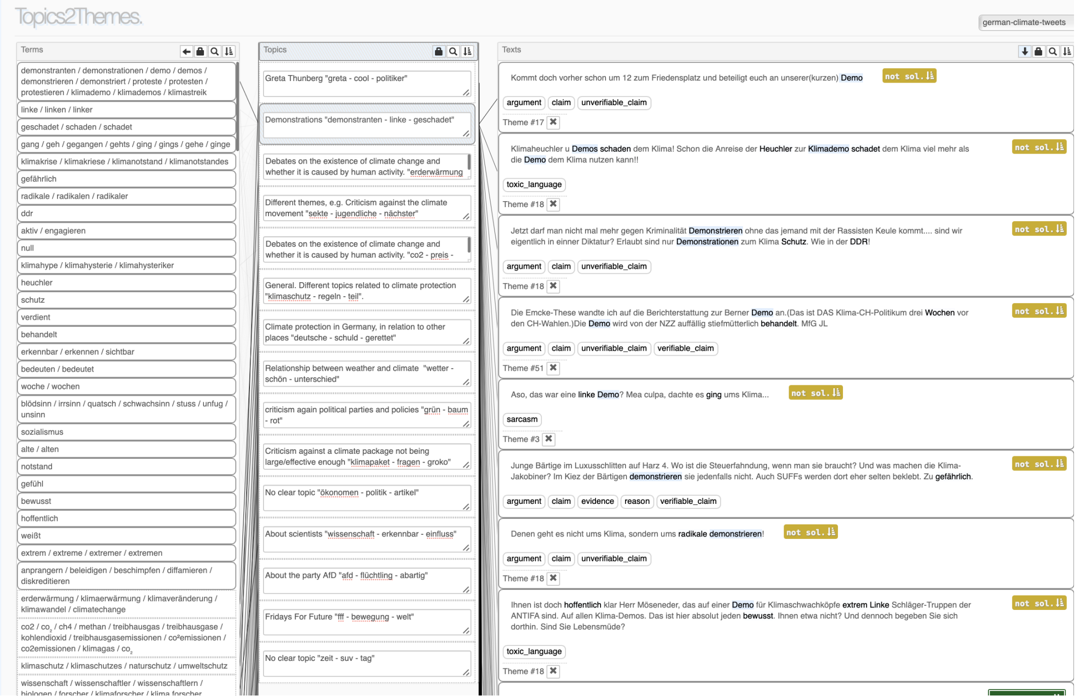
In this work, we decided to focus on one specific corpus. More specifically, we studied the content of the recently published German Climate Change Tweet Corpus (GerCCT)[1], which has been manually annotated for the existence of claims and evidence and their properties, as well as for sarcastic and toxic language (Schaefer and Stede, 2020; Schaefer and Stede, 2022). The actual content of these arguments has, so far, not been studied. We therefore aim to explore the content of the corpus, and find examples of recurring themes among the tweets. By *recurring themes* we mean recurring conceptual categories that can be identified by manually analysing the texts (Baumer et al., 2017).

The corpus contains 1,200 tweets, and we decided to set aside time to manually read a subset of these tweets in search for examples of recurring themes. In order to increase the likelihood of finding recurring content, we (i) used a topic modelling tool which automatically extracts recurring topics in a text collection, and then (ii) searched for themes in the tweets most closely associated with these automatically extracted topics. We used the topic modelling tool Topics2Themes[2] (Skeppstedt et al., 2018), which previously has been applied to extract recurring themes from a number of different types of corpora. Topics2Themes is currently being further developed and evaluated within Språkbanken Sam[3], which is a part of SWE-CLARIN.

[1] https://doi.org/10.5281/zenodo.6479492

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

[3] https://www.sprakbanken.se/sprakbankeninenglish.html



demonstranten / demonstrationen / demo / demos / demonstrieren / demonstriert / proteste / protesten / protestieren / klimademo / klimademos / klimastreik

linke / linken / linker

geschadet / schaden / schadet

gang / geh / gegangen / gehts / ging / gings / gehe / ginge

klimakrise / klimakriese / klimanotstand / klimanotstandes

gefährlich

radikale / radikalen / radikaler

ddr

aktiv / engagieren

null

klimahype / klimahysterie / klimahysteriker

heuchler

schutz

verdient

behandelt

erkennbar / erkennen / sichtbar

bedeuten / bedeutet

woche / wochen

blödsinn / irrsinn / quatsch / schwachsinn / stuss / unfug / unsinn

sozialismus

alte / alten

notstand

gefühl

bewusst

hoffentlich

weißt

extrem / extreme / extremer / extremen

anprangern / beleidigen / beschimpfen / diffamieren / diskreditieren

erderwärmung / klimaerwärmung / klimaveränderung / klimawandel / climatechange

co2 / co_2 / ch4 / methan / treibhausgas / treibhausgase / kohlendioxid / treibhausgasemissionen / co²emissionen / co2emissionen / klimagas / co_2

klimaschutz / klimaschutzes / naturschutz / umweltschutz

wissenschaft / wissenschaftler / wissenschaftlern / biologen / forscher / klimaforscher / klima forscher

Greta Thunberg "greta - cool - politiker"

Demonstrations "demonstranten - linke - geschadet"

Debates on the existence of climate change and whether it is caused by human activity. "erderwärmung

Different themes, e.g. Criticism against the climate movement "sekte - jugendliche - nächster"

Debates on the existence of climate change and whether it is caused by human activity. "co2 - preis -

General. Different topics related to climate protection "klimaschutz - regeln - teil".

Climate protection in Germany, in relation to other places "deutsche - schuld - gerettet"

Relationship between weather and climate "wetter - schön - unterschied"

criticism again political parties and policies "grün - baum - rot"

Criticism against a climate package not being large/effective enough "klimapaket - fragen - groko"

No clear topic "ökonomen - politik - artikel"

About scientists "wissenschaft - erkennbar - einfluss"

About the party AfD "afd - flüchtling - abartig"

Fridays For Future "fff - bewegung - welt"

No clear topic "zeit - suv - tag"

Kommt doch vorher schon um 12 zum Friedensplatz und beteiligt euch an unserer(kurzen) Demo

argument claim unverifiable_claim

Theme #17 ✖

Klimaheuchler u Demos schaden dem Klima! Schon die Anreise der Heuchler zur Klimademo schadet dem Klima viel mehr als die Demo dem Klima nutzen kann!!

toxic_language

Theme #18 ✖

Jetzt darf man nicht mal mehr gegen Kriminalität Demonstrieren ohne das jemand mit der Rassisten Keule kommt.... sind wir eigentlich in einner Diktatur? Erlaubt sind nur Demonstrationen zum Klima Schutz. Wie in der DDR!

argument claim unverifiable_claim

Theme #18 ✖

Die Emcke-These wandte ich auf die Berichterstattung zur Berner Demo an.(Das ist DAS Klima-CH-Politikum drei Wochen vor den CH-Wahlen.)Die Demo wird von der NZZ auffällig stiefmütterlich behandelt. MfG JL

argument claim unverifiable_claim verifiable_claim

Theme #51 ✖

Aso, das war eine linke Demo? Mea culpa, dachte es ging ums Klima...

sarcasm

Theme #3 ✖

Junge Bärtige im Luxusschlitten auf Harz 4. Wo ist die Steuerfahndung, wenn man sie braucht? Und was machen die Klima-Jakobiner? Im Kiez der Bärtigen demonstrieren sie jedenfalls nicht. Auch SUFFs werden dort eher selten beklebt. Zu gefährlich.

argument claim evidence reason verifiable_claim

Denen geht es nicht ums Klima, sondern ums radikale demonstrieren!

argument claim unverifiable_claim

Theme #18 ✖

Ihnen ist doch hoffentlich klar Herr Möseneder, das auf einer Demo für Klimaschwachköpfe extrem Linke Schläger-Truppen der ANTIFA sind. Auf allen Klima-Demos. Das ist hier absolut jeden bewusst. Ihnen etwa nicht? Und dennoch begeben Sie sich dorthin. Sind Sie Lebensmüde?

toxic_language

Theme #18 ✖

Figure 1: The three panels of Topics2Themes' user interface that show automatically extracted information. That is, the topics extracted (the mid-panel), as well as the words/word clusters (to the left) and texts (to the right) most closely associated with each one of the topics extracted.

2 Methods

2.1 The Corpus

The tweets had been collected in 2019 with the following criteria: (i) they should consist of a pair of an original *context* tweet and a *reply* to this tweet, (ii) they should be written in German, and (iii) they should contain the string "Klima" (climate). Around 12,000 tweet pairs were retrieved (Schaefer and Stede, 2020), and 1,200 of these were randomly selected for manual annotation (Schaefer and Stede, 2022). Only the reply tweet was annotated, but the context tweet was shown to the annotator to facilitate the interpretation of the reply tweet. The following annotation categories were used for annotating the replies: (i) claim (unverifiable and verifiable), (ii) evidence (reason, external evidence, internal evidence), (iii) sarcasm, (iv) toxic language. More information can be found in the two cited publications.

2.2 Automatically Extracting Topics with Topics2Themes

We applied the topic modelling tool Topics2Themes to the annotated part of the tweet pairs. The tool allows for a pre-processing of the texts, in which semantically close words are combined into one concept, and treated as one word by the topic modelling algorithm. This is primarily an automatic process, in which word2vec vectors representing the words in the corpus are clustered using dbscan clustering, and in which the created word clusters are then combined into one concept. It is, however, also possible to manually influence this process by providing manually created word clusters, as well as providing a list of words to be excluded from the clustering process. We used a word2vec model trained on German Wikipedia and German news articles for the automatic clustering.[4] We, however, also automatically

[4] https://devmount.github.io/GermanWordEmbeddings/

clustered the corpus using a model trained on tweets, and added some of the clusters from this model to the list of manually constructed clusters to use.[5] We aimed for clusters that were as large as possible (to reduce the number of features for our small corpus), but that still contained words referring to one specific concept.

There are also other configuration settings. For instance, we provided the tool with a stop word list, i.e., a list of words not to include in the topic modelling, which consisted of 2,128 words[6], as well as a list of 5 collocations. We further configured the tool to use the topic modelling algorithm non-negative matrix factorisation (Lee and Seung, 2001), to run the algorithm 500 times and only retain stably occurring topics, as well as to extract a maximum of 20 topics.[7] We also configured the tool to label the tweets with their manual annotation categories.

2.3 Manually Searching for Themes in Texts Associated with the Topics

The graphical user interface of Topics2Themes supports the user in searching for recurring themes in the texts extracted by the topic model. The automatically extracted output of the topic modelling algorithm – i.e., (i) the list of topics extracted, (ii) words/word clusters closely associated with the topics, and (iii) texts closely associated with the topics – are displayed in three different panels in the user interface. There is also a forth panel, to which the user can manually add themes identified in the texts. While the user creates themes and assigns them to the texts, the Topics2Themes tool continuously re-trains a Nearest Neighbour classifier on the text-theme associations. Thereby, by re-sorting the list of previously created themes, the tool can automatically provide the user with suggestions for themes matching the text that is being analysed. Since the model is being re-trained for each new text-theme association created, it has a chance to continuously improve and adapt to the user's theme-assignment strategy.

We used this theme-assignment functionality of the graphical user interface of Topics2Themes. That is, we searched for recurring themes among the tweets most closely associated with the automatically extracted topics, and manually added them to the forth panel of the tool.

3 Results and Discussion

Topics2Themes extracted 15 topics, as shown in the *Topics* panel of its user interface (Figure 1). The topic descriptions were written after skimming the texts most closely associated with each one of these 15 topics. The topics often had quite a diverse content. However, two topics had many closely associated texts that we described as containing "debates on the existence of climate change and whether it is caused by human activity". Other examples of groups of similar tweets associated with one topic were tweets on Greta Thunberg, on demonstrations, on the difference between climate and weather, on scientists, on the German party AfD, and on Fridays for Future. The extracted topics typically relied more on one single word cluster than on word co-occurrences. E.g., the eight texts most closely associated with the "demonstrations" topic, selected in Figure 1, all contain words from the "demonstration/protest/climate strike" word cluster. However, only two of the texts contain words from the second most closely associated word cluster (variants of "left-wing"). This is probably an effect of the topic modelling algorithm being applied to short texts, which provide a small statistical basis for finding word co-occurrences. This might also explain the diverse content of some of the topics.

For each topic, we read the eight most closely associated tweets (i.e., around 10% of the corpus) in search for recurring themes on a level more granular than the topic level. We identified 53 potential themes in these texts. Most of the theme candidates only occurred once, but five themes occurred twice and 14 themes occurred at least three times (see Table 1).

When reading the tweets, we noticed that the meaning of some of them (21 tweets) could not be interpreted, probably due to lack of context. As a next step, we therefore plan to not only include the reply

[5] https://www.cl.uni-heidelberg.de/english/research/downloads/resource_pages/GermanTwitterEmbeddings/GermanTwitterEmbeddings_data.shtml

[6] We started with the list provided here https://github.com/solariz/german_stopwords and then manually extended it with words extracted by the topic modelling algorithm.

[7] Configuration files can be found at: https://github.com/mariask2/Climate-tweets-in-german.

Theme
(a) Criticism against a climate package not being large/effective enough, or demands for a better one.
(b) Argumentation claiming that it is not proven that human activity causes climate change, that climate change is not a problem, or that we cannot know anything about future climate.
(c) Argumentation stating that climate change exists, is caused by human activity and is a problem.
(d) Criticism against Greta Thunberg, e.g. claims that she is being forced by others to protest.
(e) Criticism against the field and background of a scientist in relation to the subject of climate change.
(f) Criticism of that climate change receives too much attention/too much is done/there's a hysteria.
(g) Criticism of the climate movement, e.g., for being a sect, searching for street conflicts, too fanatic.
(h) Critique against climate demonstrations, e.g., claims that it is not allowed to demonstrate against anything else, more interest in radical demonstration than climate.
(i) Critique against party strategies in relation to climate change.
(j) Criticism against those working against climate change doing activities that emit a lot of CO2.
(k) Discussions about the relation between weather and climate.
(l) Critique against that (or trying to find reasons why) someone does not worry about climate change.
(m) Criticism towards politicians for not doing enough for the climate.
(n) Discussions about different economic instruments for climate protection.

Table 1: Themes found in at least three tweets.

tweet in the experiment, but also the original context tweet. We also plan to apply the topic modelling tool to the entire corpus collected, which contains 12,000 tweet pairs.

Acknowledgements

The work presented here was funded by the Vinnova project "Using the SB Sam NLP tools for manual and automatic annotation of climate change texts" (Ref no: 2021-03973) and "HUMINFRA: National infrastructure for research in the humanities and social sciences (Ref no: 2021-00176)". The development of Topics2Themes is funded by the Swedish Research Council (most recently, ref no: 2017-00626).

References

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410, June.

IPCC. 2021. Climate Change 2021, The Physical Science Basis, Summary for Policymakers. Available from https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_SPM_final.pdf.

Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556 – 562.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online, December. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2022. GerCCT: An annotated corpus for mining arguments in German tweets on climate change. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6121–6130, Marseille, France, June. European Language Resources Association.

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.

Linguistic Framing of Political Terror: Distant and Close Readings of the Discourse on Terrorism in the Swedish Parliament 1993–2018

Magnus P. Ängsal
University of Gothenburg, Sweden
magnus.pettersson.angsal@sprak.gu.se

Daniel Brodén
University of Gothenburg, Sweden
daniel.broden@gu.se

Mats Fridlund
University of Gothenburg, Sweden
mats.fridlund@gu.se

Leif-Jöran Olsson
University of Gothenburg, Sweden
leif-joran.olsson@gu.se

Patrik Öhberg
University of Gothenburg, Sweden
patrik.ohberg@gu.se

Abstract

This paper provides a study of the discourse on terrorism in Swedish parliamentary debate 1993–2018. The aim is to explore how terrorism is discursively constructed in parliamentary deliberations, drawing on the resources of Swe-Clarin in the form of the corpus tool Korp and the linguistic concept of 'frame'. To map meanings attached to terrorism we pursue two research questions: what framing elements are connected to 'terrorism' and 'terrorist' in parliamentary speeches as 1) simplexes and 2) as part of compounds along the lines of controversies and party affiliations? The latter research question is probed through distant and close readings of the specific compound *statsterrorism* ('state terrorism'). Our findings show that terrorism is typically framed as located outside of Sweden and as tied to Islamism, but the question of what countries are associated with state terrorism depends on the political affiliation of the interlocutor. The compound *statsterrorism* is most prominently used by the left and green parties and then commonly associated with Israel and Turkey. We conclude by suggesting that a widened inquiry into compounds, in general as well as diachronically, is likely a productive way of expanding the scope of our research.

1 Introduction

This study provides a digital history of Swedish terrorism by means of a combination of distant and close readings of Swedish parliamentary debate speeches 1993–2018. The investigation draws on the resources of Swe-Clarin in the form of Korp, the corpus infrastructure of Språkbanken Text, providing the parliamentary corpus and the corpus tool used for the analysis (c.f. Borin, Forsberg and Roxendal 2012).

The paper builds on prior interdisciplinary studies tied to Swe-Clarin's outreach activities, combining humanities and language technology scholarship to explore the discourse on terrorism in Sweden (see e.g. Fridlund et al., 2020). Specifically, we ask two main research questions regarding the parliamentary understanding of the phenomenon of terrorism during the period in focus: (1) What framing elements understood as discursive key meanings are connected to the words *terrorism* and *terrorist* (same spelling in Swedish and English) when used in parliamentary speeches as simplexes, and (2) as part of compounds along the lines of controversies and party affiliations? The second research question is addressed by means of a close reading of the compound *statsterrorism* ('state terrorism').

The parliamentary debate data studied belongs to the broader Swedish political discourse on terrorism which is explored further in the Swe-Clarin supported *SweTerror*-project that this study is a part of

(Edlund et al. 2022). Here, to map the meanings attached to terrorism in the data, we use Korp for searches and queries based on automatic linguistic annotations with structured result presentations: a so called KWIC (keyword in context) contextual hit list; statistical data of keyword occurrences in subcorpora 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, e.g. typical subjects and objects of a verb, and nominal premodifiers (e.g. adjectives) and post-modifiers (prepositional phrases or main verbs of relative clauses). The word picture can be used as a topical map to guide users to closer readings of the corpus. Korp also supports navigation between the statistics, trend-graph and word picture views, and the KWIC view allows close reading of individual hits in their specific close context.

Notably, despite its prevalent use, there is no generally accepted definition of terrorism and it has been described as an 'essentially contested concept' (Weinberg et al., 2004). The theoretical concept of discourse guiding our readings adheres to a notion of discourse as a virtual collection of intertextually entwined utterances relating to a macro-topic (here: terrorism) as a social practice, thereby signifying, reinforcing and questioning societal structures of power and agency (cf. Reisigl and Wodak 2009). In order to scrutinize conceptualizations of terrorism understood as discursive formations, we utilize the analytical concept of frame (cf. Entman 1993), meaning conceptual frames of understanding and interpretation in which concrete linguistic acts are always embedded.

2 Analysis of the Simplexes *Terrorism* and *Terrorist*

The search string *terrorism* generates 4.755 hits in the Korp parliamentary corpus, the vast majority of the instances, 4.399, being different grammatical forms of the simplex *terrorism*, whereas the rest (356 hits) are complexes with *terrorism* or *terrorist* either as the modifier or head constituent. One striking finding is that terrorism as a focal point of political debate is almost absent before 2001, another one being the rise in number of hits in 2015. The former result can be explained by the attacks in the US on September 11, 2001, while the latter can likely be attributed to the rise of the Islamic State (Daesh, ISIS), which was followed by numerous violent acts in Western European cities, including the 2017 lethal lorry attack on Drottninggatan in downtown Stockholm.

Turning to KWIC and word pictures for *terrorism*, the search run with Korp displays notable results. The strongest and, in absolute numbers, most frequent collocate of *terrorism* is the preposition *mot* ('against'), which indicates discursive key meaning adhering to the word: terrorism is typically conceived of as a phenomenon to overcome or combat. The finding aligns with the strongest collocate among verbs, namely *bekämpa* ('combat'). Turning to the premodifiers found in the word picture, *internationell, islamistisk, storskalig* ('large-scale'), *global* and *jihadistisk/a* (the latter represented in two different grammatical forms, indefinite and definite) are the strongest collocates. It discloses that the frame invoked by *terrorism* in this data entails an agent slot that is occupied primarily by Islamist terrorism. One example from the data illustrates how terrorism is framed as a Jihadist phenomenon. In this example, voiced by, then, right-wing Member of Parliament Mikael Jansson (Sweden Democrats), Islamist terrorism is pictured as a threat: *'Den islamistiska terrorismen bör vi européer möta gemensamt.* [We Europeans ought to jointly encounter Islamist terrorism.]' (2016-05-19)

The study also encompasses an inquiry into the personal noun *terrorist* and its contexts in the corpus. The intriguing question here is how terrorists are characterized, i.e. how they are framed in the data at hand. The total number of uses of the word *terrorist* in the corpus amounts to 3.360 over the given period. However, of these hits 1.996 are compounds, whereas only 1.364 instances are made up by simplexes. This can certainly be attributed to the fact that *terrorist-* is a more productive word form in compounds than *terrorism-*. As is the case with *terrorism*, the word *mot* ('against') is the single strongest (although not most frequent) preposition preceding the term in the parliamentary speeches. It should be noted, however, that also *för* ('for' or 'in favor of') ranks high (second position) among the strongest prepositions. This circumstance calls for further close analysis of the contextual embeddings. Among the premodifiers, the strongest collocate is *islamistisk* ('Islamist'), followed by *palestinsk* ('Palestinian') and *potentiell* ('potential'). Nevertheless, *internationell*, which was the strongest collocate for *terrorism*, is absent. Notably, the terrorist – unless s/he is a potential agent occurring in acts of denominating without referring specifically to certain terrorists – is more clearly situated within a specific national or political context. This conclusion might be supported also by the fact that the single strongest collocate among postmodifiers is the prepositional phrase *'i Israel'* ('in Israel'). The terrorist frame, then, locates the agent in a context shaped by long-existing conflicts as well as by more recent forms of Islamism.

As far as the predicate verbs preceding and succeeding the word *terrorist* are concerned, there are lexemes to be found that indicate both a critical attitude towards terrorists (*jaga* 'hunt'; *bekämpa* 'combat') and, as it seems, endorsement (*belöna* 'reward'; *hylla* 'praise'). This could make the case for a differentiated frame, or possibly for the existence of different frames, with respect to the attitudes conveyed towards terrorists. Close reading review of the instances of *hylla* in the narrow context of *terrorist*, however, discloses an entanglement between these two lexical units of a different kind. In most instances, *hylla* is used in the context of Israel/Palestine. The speaker thereby typically accuses someone else – most often the West Bank Palestinian authorities – of praising terrorists.

3 Analysis of Compounds with *Terrorism*

As far as compounds with *terrorism* is concerned, there are seven lexemes with at least 10 instances in the corpus: *terrorismbekämpning* (*-bekämpning* '-combating'), 82 instances; *statsterrorism* (*stats-* 'state-), 63 instances; *terrorismresa* (*-resa* '-travel'), 38 instances; *terrorismlagstiftning* (*-lagstiftning* '-legislation'), 31 instances; *terrorismområdet* (*-området* '-area'), 10 instances; *terrorismrelaterad* (*-relaterad* '-related'), 10 instances; *terrorismsyfte* (*-syfte* '-purpose'), 10 instances.

Turning to the usage of compounds over time, there is a moderate increase around 2002, and comparatively many hits can be found at the end of the period, with distinct peaks in the years from 2015 until 2018. The first rise in 2002, with 26 individual occurrences, is to be explained by the attacks in the US on September 11th the year before. From 2015 and onwards, slightly more than half of all instances of compounds, 182 hits, occur. This circumstance can most certainly be attributed to an increased focus on terrorism as a political issue of great societal urgence, due to the emergence of the Islamic State in the Middle East. Indicative in this respect is the sudden appearance of the compound *terrorismresa* ('terrorism travel'), with, all in all, 38 instances, in the year of 2015.

The paper also entails a close reading of the compound *statsterrorism* in relation to the party affiliations of the Members of Parliament. *Statsterrorism* is, in fact, a somewhat curious compound in the context of terrorism, since it denotes an atypical form of terrorism (cf. Jackson 2011), namely terrorist actions taken by states, and not by clandestine groups or movements opposing a state apparatus. There are 63 tokens of *statsterrorism* in the data. The intriguing questions are, first, what the points of reference are for 'state terrorism', i.e., who the alleged agent is, and second, how these assignments to various states and other organizational agents can be considered against the backdrop of party affiliation.

Israel (16 hits) is the most common agent assigned to state terrorism in the corpus, followed by Iran (15), Turkey (11) and Russia (9). Among the agents mentioned, there is thus no designation of Sweden as a state-terrorist nor of any European or North American state. This differs from what is known from the record of parliamentary speeches 1968-1970 when the USA was often described as using terror during the Vietnam war (cf. Fridlund et al. 2022). However, a 'Western' state in the form of Israel ranks highest with 16 hits, whereas Iran is represented with 15 instances, followed by Turkey and Russia. Among those, especially Russia is well-known from the historical record as an alleged state-terrorist. Russian state terror has even been recognized as the initial provocation of the first emergence of substate terrorism in the 1870s Russia (cf. Miller 2013).

Third and last, the distribution has a bias towards the Left-Right political paradigm in so far as Israel and Turkey only occur among the three left-leaning parties, the Green, Left and Social Democratic parties, while Iran and Russia appear across the political spectrum. Iran, furthermore, is the most common agent for state terrorism in Social Democratic contributions.

4 Conclusion

This paper provides an exploratory attempt to better understand how a political concept – terrorism – has been constituted linguistically in Swedish politics. Drawing on a mixed-methods approach to the parliamentary debate during the period 1993–2018, we have explored elements of the terrorism frame, understood as discursive meanings connected to the use of the words *terrorism* and *terrorist* as simplexes or as parts of compounds in parliamentary speeches.

Our study tentatively indicates that in the time-period in focus, the agents behind terrorism are often framed as Islamists, or – more generally and more underspecified – as international actors. To some extent, however, the agent slot can be found enriched with states as agents, which is substantiated in the minor case-study on the compound *statsterrorism*. Furthermore, we highlight the significance of party affiliations and lines of controversies. The very occurrence of the word *statsterrorism* is more frequent in speeches delivered by MPs from the left and green parties, typically, however not in all instances, designating Israel or Turkey as agents. Furthermore, by applying the word picture function in the corpus tool Korp further proof is substantiated for the notion of terrorism as something negative that one needs to overcome. As far as the terrorist frame is concerned, similar results can be found.

Notable is also that terrorism is rarely debated in our material 1993–2001, which indicates that terrorism was a marginal parliamentary political issue during this period. This is not to say that political violence resembling or being de facto identical with terrorist acts were not discussed. However, if such debates occurred, they were not centered around the concept of terrorism. Moreover, distinct quantitative peaks can be detected 2015–2018. This is likely due to an increased focus on the movement known as the Islamic State concerning violent deeds in Europe and on Swedish citizens joining IS in Syria and Iraq. The latter can particularly be substantiated by reviewing the hits of the complex lexeme *terrorismresa* ('terrorism travel'), which figures exclusively during this period.

On another level, the paper constitutes an explorative methodological study of the possibility of discerning semantic frame elements combining distant and close reading approaches to the corpus of Parliamentary speeches in Korp. As expected, our findings confirm the potential of combining computationally driven and interpretative approaches for analysing the discursive construction of terrorism, as well as other issues, in the Swedish Parliament. Further inquiries into compounds would widen and deepen the research scope. Another point of advancement with respect to further scrutiny is to systematically apply a diachronic perspective on the data. Such an expansion would allow for mapping out frequencies and trends in the usage of specific lexemes with *terrorism*- and *terrorist*- over time.

References

Borin, L., Forsberg, M. and Roxendal, J. 2012. Korp – the corpus infrastructure of Språkbanken. Calzolari, N. et al. (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC' 12)*, 474–478.

Edlund, J., Brodén, D., Fridlund, M., Lindhé, C., Olsson, L-J., Ängsal, M.P. and Öhberg, P. 2022. A multimodal digital humanities study of terrorism in Swedish politics: an interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018. Arai, K. (Ed.), *Intelligent Systems and Applications: IntelliSys 2021.* Springer, Cham, 435–449.

Entman, R. 1993. Framing: toward clarification of a fractured program. *Journal of Communication* 43(4):51–58.

Fridlund, M., Brodén, D., Olsson, L-J. and Ängsal, M.P. 2022. Codifying the debates of the Riksdag: towards a framework for semi-automatic annotation of Swedish parliamentary discourse. La Mela, M., Norén, F. and Hyvönen, E. (Eds.), *DiPaDa 2022: Proceedings of the Digital Parliamentary Data in Action (DiPaDa 2022) Workshop*, 167–175.

Fridlund, M., Olsson, L-J., Brodén, D. and Borin, L. (2020). Trawling the Gulf of Bothnia of news: a big data analysis of the emergence of terrorism in Swedish and Finnish newspapers, 1780–1926. C. Navarretta and Eskevich, M. (Eds.) *CLARIN Annual Conference Proceedings 2020*, 61–65.

Jackson, R. 2011. In defence of 'terrorism': finding a way through a forest of misconceptions. *Behavioral Sciences of Terrorism and Political Aggression* 3(2):116–130.

Miller, M.A. 2013. *The Foundations of Modern Terrorism: State, Society, and the Dynamics of Political Violence*. Cambridge University Press, Cambridge, UK.

Reisigl, M. and Wodak, R. 2009. The discourse-historical approach (DHA). Wodak, R. and Meyer, M. (Eds), *Methods of Critical Discourse Analysis*, 2nd revised edition. Sage Pub, London, Thousand Oaks, 87–121.

Weinberg, L., Pedahzur, A. and Hirsch-Hoefler, S. (2004). The challenges of conceptualizing terrorism, *Terrorism and Policical Violence* 16(4):777–794.