

CLARIN Annual Conference Proceedings

2019

Edited by

Kiril Simov, Maria Eskevich

30 September – 2 October 2019
Leipzig, Germany

Please cite as:
CLARIN Annual Conference Proceedings, 2019. ISSN 2773-2177 (online).
Eds. K. Simov and M. Eskevich.
Leipzig, Germany: CLARIN, 2019.

Programme Committee

Chair:

- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)

Members:

- Lars Borin, Språkbanken, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Griet Depoorter, Institute for the Dutch Language (NL/Vlanders)
- Jens Edlund, KTH Royal Institute of Technology (SE)
- Roald Eiselen, South African Centre for Digital Language Resources (ZA)
- Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)
- Karlheinz Mörth, Institute for Corpus Linguistics and Text Technology, 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)
- Eiríkur Rögnvaldsson, University of Iceland (IS)
- Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia & Tilde (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadić, University of Zagreb (HR)
- Jurgita Vaičėnienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Ilze Auziņa, LV
- Jaco Badenhorst, ZA
- Andrea Bellandi, IT
- Lars Borin, SE
- Sonja Bosch, ZA
- Daan Broeder, NL
- Karen Calteaux, ZA
- Gregory Crane, DE
- Roberts Dargis, LV
- Johannes Dellert, DE
- Griet Depoorter, NL/Vlanders
- Roald Eiselen, ZA
- Tomaž Erjavec, SI
- Darja Fišer, SI
- Francesca Frontini, FR
- Dimitrios Galanis, GR
- Maria Gavrilidou, GR
- Thomas Gloning, DE
- Luis Gomes, PT
- Mikus Grasmanis, LV
- Riccardo Del Gratta, IT
- Eva Hajičová, CZ
- Erhard Hinrichs, DE
- Martin Holub, CZ
- Neeme Kahusk, EE
- Pawel Kamocki, DE
- Fahad Khan, IT
- Nicolas Larrousse, FR
- Laska Laskova, BG
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Valeria Quochi, IT
- Jan Odijk, NL
- Petya Osenova, BG
- Christophe Parisse, FR
- Maciej Piasecki, PL
- Stelios Piperidis, GR
- Hannes Pirker, AT
- Martin Puttkammer, ZA
- Eirikur Rögnvaldsson, IS
- Magda Ševčíková, CZ
- Marko Robnik Šikonja, SI
- João Silva, PT
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Pavel Stranak, CZ
- Christian Thomas, DE
- Mateja Jemec Tomazin, SI
- Thorsten Trippel, DE
- Andrius Utkā, LT
- Jurgita Vaičėnonienė, LT
- Darinka Verdonik, SI
- Jernej Vičič, SI
- Kadri Vider, EE
- Martin Wynne, UK
- Tanja Wissik, AT

CLARIN 2019 submissions, review process and acceptance

- Call for abstracts: 22 January 2019, 25 February 2019
- Submission deadline: 29 April 2019
- In total 56 submissions were received and reviewed (three reviews per submission)
- Face-to-face PC meeting in Riga: 18-19 June 2019
- Notifications to authors: 1 July 2019
- 44 accepted submissions: 25 oral presentations, 19 posters/demos

More details on the paper selection procedure and the conference can be found at <https://www.clarin.eu/event/2019/clarin-annual-conference-2019-leipzig-germany>.

Table of Contents

Thematic Session: Humanities and Social Science research enabled by language resources and technology

<i>Named Entity Annotation for Ancient Greek with INCEPTION</i>	
Monica Berti	1
<i>Enriching Lexicographical Data for Lesser Resourced Languages: A Use Case</i>	
Dirk Goldhahn, Thomas Eckart and Sonja Bosch	5
<i>CLARIN-Supported Research on Modification Potential in Dutch First Language Acquisition</i>	
Jan Odijk	9

Parallel Session 1: Design and Construction of the CLARIN Infrastructure

<i>Training Workshops in the Bi-directional Model of the Language Technology Infrastructure Development</i>	
Maciej Piasecki and Jan Wiczorek	14
<i>OpeNER and PANACEA: Web Services for the CLARIN Research Infrastructure</i>	
Davide Albanesi and Riccardo Del Gratta	19
<i>CLARIAH Chaining Search: A Platform for Combined Exploitation of Multiple Linguistic Resources</i>	
Peter Dekker, Mathieu Fannee and Jesse De Does	24

Parallel Session 2: Use of the CLARIN Infrastructure

<i>Manually PoS Tagged Corpora in the CLARIN Infrastructure</i>	
Tomaž Erjavec, Jakob Lenardič and Darja Fišer	28
<i>A Use Case for Open Linguistic Research Data in the CLARIN Infrastructure. The Open Access Database for Adjective-Adverb Interfaces in Romance</i>	
Gerlinde Schneider, Christopher Pollin, Katharina Gerhalter and Martin Hummel	32
<i>CLARIN Web Services for TEI-annotated Transcripts of Spoken Language</i>	
Bernhard Fisseni and Thomas Schmidt	36

Parallel Session 3: Use of the CLARIN Infrastructure

<i>Using DiaCollo for Historical Research</i>	
Bryan Jurish and Maret Nieländer	40
<i>Corpus-Preparation with WebLicht for Machine-made Annotations of Examples in Philosophical Texts</i>	
Christian Lück	44
<i>Lifespan Change and Style Shift in the Icelandic Gigaword Corpus</i>	
Lilja Björk Stefánsdóttir and Anton Karl Ingason	48
<i>Studying Disability Related Terms with Swe-CLARIN Resources</i>	
Lars Ahrenberg, Henrik Danielsson, Staffan Bengtsson, Hampus Arvå Lotta Holme and Arne Jönsson	
52	

Parallel Session 4: Legal Issues

<i>To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest</i>	
Krister Lindén, Aleksei Kelli and Alexandros Nousias	56
<i>Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection</i>	
Inga Kaija and Ilze Auziņa	61
<i>Liability of CLARIN Centres as Service Providers: What Changes with the New Directive on Copyright in the Digital Single Market?</i>	
Pawel Kamocki, Andreas Witt, Erik Ketzan and Julia Wildgans	65
<i>The Extent of Legal Control over Language Data: the Case of Language Technologies</i>	
Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits and Age Värvi	69

Parallel Session 5: CLARIN in Relation to Other Infrastructures and Applications

<i>User Support for the Digital Humanities</i>	
Tommi A Pirinen, Hanna Hedeland and Heidemarie Sambale	75
<i>CLARIN AAI and DARIAH AAI Interoperability</i>	
Peter Gietz and Martin Haase	80
<i>Word at a Glance – a Customizable Word Profile Aggregator</i>	
Tomáš Machálek	85
<i>Technical Solutions for Reproducible Research</i>	
Alexander König and Egon W. Stemle	89

Parallel Session 6: CLARIN in Relation to Other Infrastructures and Metadata

<i>Approaches to Sustainable Process Metadata</i>	
Kerstin Jung and Markus Gärtner	93
<i>The Best of Three Worlds: Mutual Enhancement of Corpora of Dramatic Texts (GerDraCor, German Text Archive, TextGrid Repository)</i>	
Frank Fischer, Susanne Haaf and Marius Hug	97
<i>Mapping METS and Dublin Core to CMDI: Making Textbooks Available in the CLARIN VLO</i>	
Francesca Fallucchi and Ernesto William De Luca	104
<i>What Got Connected - Parthenos Ending</i>	
Matej Durco, Klaus Illmayer and Stefan Resch	108

Poster session

<i>A New Gold Standard for Swedish NERC</i>	
Lars Ahrenberg, Leif-Jöran Olsson and Johan Frid	112
<i>From OCR to Digital Editions</i>	
Saranya Balasubramanian	116
<i>Enhancing Lexicography by Means of the Linked Data Paradigm: LexO for CLARIN</i>	
Andrea Bellandi, Fahad Khan and Monica Monachini	120
<i>Aggregating Resources in CLARIN: FAIR Corpora of Historical Newspapers in the German Text Archive</i>	
Matthias Boenig and Susanne Haaf	124
<i>CLARIN and Digital Humanities. A Successful Integration.</i>	
Elisabeth Burr, Marie Annisius and Ulrike Fußbahn	129
<i>AcTo: How to Build a Network of Integrated Projects for Medieval Occitan</i>	
Gilda Caïti-Russo, Jean-Baptiste Camps, Gilles Couffignal, Francesca Frontini, Hervé Lieutard, Elisabeth Reichle and Maria Selig	134
<i>A Parsing Pipeline for Icelandic Based on the IcePaHC Corpus</i>	
Tinna Frímann Jökulsdóttir, Anton Karl Ingason and Einar Freyr Sigurðsson	138
<i>Optimizing Interoperability of Language Resources with the Upcoming IIF AV Specifications</i>	
Jochen Graf, Felix Rau and Jonathan Blumtritt	142
<i>SpeCT 2.0 – Speech Corpus Toolkit for Praat</i>	
Mietta Lennes	147
<i>CLARIN-IT and the Definition of a Digital Critical Edition for Ancient Greek Poetry: a New Project for Ancient Fragmentary Texts with a Complex Tradition</i>	
Anika Nicolosi, Monica Monachini and Beatrice Nava	150
<i>Research Data of a PhD Thesis Project in the CLARIN-D Infrastructure. “Texts of the First Women’s Movement” / “Texte der ersten Frauenbewegung (TdeF)” as Part of the German Text Archive</i>	
Anna Pfundt, Melanie Grumt Suárez and Thomas Gloning	155
<i>Granularity versus Dispersion in the Dutch Diachronical Database of Lexical Frequencies TICCLAT</i>	
Martin Reynaert, Patrick Bos and Janneke van der Zwaan	159
<i>Cross Disciplinary Overtures with Interview Data: Integrating Digital Practices and Tools in the Scholarly Workflow.</i>	
Stefania Scagliola, Louise Corti, Silvia Calamai, Norah Karrouche, Jeannine Beeken, Arjan van Hessen, Christoph Draxler, Henk van den Heuvel and Max Broekhuizen	163
<i>The Rise of the Definiteness Effect in Icelandic</i>	
Einar Freyr Sigurðsson and Anton Karl Ingason	167
<i>Integrated Language and Knowledge Resources for CLaDA-BG</i>	
Kiril Simov and Petya Osenova	171
<i>Application of a Topic Model Visualisation Tool to a Second Language</i>	
Maria Skeppstedt, Magnus Ahltop, Andreas Kerren, Rafal Rzepka, Kenji Araki	176
<i>CTS-R: Connecting Canonical Text Services with the Statistical Analytics Environment R</i>	
Jochen Tiepmar	180
<i>Shapeshifting Digital Language Resources - Dissemination Services on ARCHE</i>	
Martina Trognitz and Matej Ďurčo	184
<i>Wablieft: An Easy-to-Read Newspaper corpus for Dutch</i>	
Vincent Vandeghinste, Bram Bulté and Liesbeth Augustinus	188

Named Entity Annotation for Ancient Greek with INCEpTION

Monica Berti

Institute of Computer Science – Digital Humanities
University of Leipzig, Germany
monica.berti@uni-leipzig.de

Abstract

This paper presents experimental work on Named Entity Recognition and Annotation for ancient Greek using INCEpTION, a web-based annotation platform built on the CLARIN tool WebAnno. Data described in the paper is extracted from the *Deipnosophists* of Athenaeus of Naucratis.

1 Introduction

This paper describes preliminary work for semi-automatic recognition and annotation of named entities (NEs) in ancient Greek. High quality annotations of NEs in this language are still missing and data presented in this paper has been extracted from the text of the *Deipnosophists* of Athenaeus of Naucratis (2nd-3rd century CE). Annotations are visualized, disambiguated, and linked with INCEpTION, which is a web-based platform built on WebAnno that was implemented for the CLARIN community.

2 Named Entity Recognition and Annotation for Ancient Greek

Named Entity Recognition (NER) is a relatively mature technology in Natural Language Processing (NLP) that is generating a lot of interest to the communities of Digital Humanities and historical languages (van Hooland et al., 2015; Nouvel et al., 2016). Concerning Classical antiquity, experiments have been carried out for Latin (Erdmann et al., 2017) and for ancient Greek in the beta version of Trismegistos People (Broux and Depauw, 2015).¹ NER is also in the agenda of the Classical Language Toolkit (Burns, 2019).² In spite of that and even if linguistic annotations of ancient Greek and Latin texts are growing (Celano, 2019), high quality annotations of NEs in these languages are still missing.³

3 Named Entity Extraction in Athenaeus of Naucratis

In order to extract and annotate NEs in ancient Greek, preliminary work has been accomplished with the text of the *Deipnosophists* of Athenaeus of Naucratis (2nd-3rd century CE), which is a rich collection of proper names pertaining to a wide variety of typologies like personal names, peoples, places, groups, languages, festivals, astronomical and meteorological phenomena, chronological data and currencies. Athenaeus' work is also a huge mine of references to more than 900 authors of Classical literature and their writings (Braund and Wilkins, 2000; Berti et al., 2016).

NER has been performed on the text of the Teubner edition of the *Deipnosophists* by Georg Kaibel, which contains 264,750 tokens distributed in 15 books for a total of 1,328 paragraphs and 21,460 sentences. In order to identify NEs, the text has been tokenized, capitalized words have been extracted,

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

¹<https://www.trismegistos.org/ref/index.php>

²<http://cltk.org>

³The open source annotation tool Recogito (<https://recogito.pelagios.org>) provides automatic NER tagging for historical data using Stanford CoreNLP (English, French, German and Spanish), experimental Latin NER with the Herodotus Latin NER plugin (Erdmann et al., 2017), and experimental Hebrew NER with the Kima NER plugin (<https://geo-kima.org>).

and non-relevant words have been removed.⁴ Extracted words have been lemmatized by querying Morpheus as well as a collection of lemmatized ancient Greek texts.⁵ Missing lemmata (ca. 50%) have been manually added and checked through Logeion⁶ and the online Thesaurus Linguae Graecae (TLG).⁷

As of early 2019, the result is the extraction of 22,924 inflected forms of single NEs corresponding to 8,435 unique forms and 4,470 lemmata. Lemmata have been used to query external authority lists to obtain a first provisional set of disambiguated annotations. Annotations of personal and geographical names have been automatically generated by querying lemmata in the online Lexicon of Greek Personal Names (LGPN) (ca. 63% of all NEs)⁸ and in the Pleiades gazetteer (ca. 17% of all NEs).⁹

Missing entities have been manually annotated. For the annotation, tags commonly used in computational linguistics for labeling generic NE types have been adopted: LOC, LOCderiv, ORG, ORGderiv, OTH, PER, PERderiv (Nouvel et al., 2016) (see Table 1). Data is stored in an SQL database, whose entries can be publicly interrogated in the *Named Entities Digger* and in the *Named Entities Concordance* of the Digital Athenaeus project.¹⁰ Every occurrence of each NE is identified with a URN according to the Canonical Text Services (CTS) protocol (Berti et al., 2016). For example, Πλάτωνος [Platonos] in Ath., *Deipn.* 9.37 is identified as `urn:cts:greekLit:tlg0008.tlg001.perseus-grc2:9.37@πλάτωνος[1]`.¹¹

NE Class	Semantic sub-classes	Number of occurrences
LOC	<i>cities, regions, islands, mountains, rivers, etc.</i>	2,151
LOCderiv	<i>location_deriv</i>	4,377
ORG	<i>festivals and Panhellenic games</i>	129
ORGderiv	<i>organization_deriv</i>	17
OTH	<i>works, months, constellations, currencies, languages, groups, etc.</i>	1,916
PER	<i>gods, persons, personifications, authors, etc.</i>	14,043
PERderiv	<i>person_deriv</i>	291
Total		22,924

Table 1: Named Entities in the *Deipnosophists*.

4 Annotating Ancient Greek Named Entities with INCEPTION

Pre-annotated data described in section 3 has been imported into INCEPTION¹² (Klie et al., 2018), a web-based platform for semantic text annotation. The INCEPTION platform re-uses parts of WebAnno¹³ (Eckart de Castilho et al., 2016), a text annotation tool that was originally developed for the CLARIN community but that at the time of writing is mainly maintained as part of the INCEPTION project. INCEPTION combines WebAnno with new functionalities such as the knowledge management, search, and more. Both tools support a wide range of text annotation tasks including NE annotation. Moreover, INCEPTION supports interactive and semantic annotation as for example concept linking, fact linking, and knowledge base population.

⁴Modern editions of ancient Greek sources generally capitalize words corresponding to proper names and words after a full stop. In the edition of the *Deipnosophists* by Kaibel only the first word of the first paragraph of each book has been capitalized for a total of 15 occurrences, while beginning words of other paragraphs are typed in lower case unless they are proper names. Sometimes Kaibel prints content-related words entirely in capital letters and they have been identified and removed. See Nouvel et al. (2016), p. 28 on the use of capitalization as a method for extracting proper names.

⁵<https://github.com/gcelano/LemmatizedAncientGreekXML>

⁶<http://logeion.uchicago.edu>

⁷<http://stephanus.tlg.uci.edu>

⁸<http://www.lgpn.ox.ac.uk>

⁹<https://pleiades.stoa.org>

¹⁰<http://www.digitalatheneus.org>

¹¹This CTS URN means that this is the first occurrence of the word Πλάτωνος in Ath., *Deipn.* 9.37 (ed. Kaibel, Perseus XML file). On the Perseus Catalog and related data repositories, see (Babeu, 2019).

¹²<https://inception-project.github.io>

¹³<https://webanno.github.io>

INCEpTION is currently used to visualize, correct, and nest annotations of single ancient Greek NEs that have been semi-automatically extracted with the method described in section 3. Data has been imported into INCEpTION as TSV files generated according to the WebAnno TSV 3.2 file format.¹⁴ Each file includes the text of single paragraphs of the *Deipnosophists* with sentences split in separate lines.¹⁵ Figure 1 shows how pre-annotated data is visualized in INCEpTION in separate numbered lines. Single entities are annotated with two layers (*Named entity* and *lemma*).

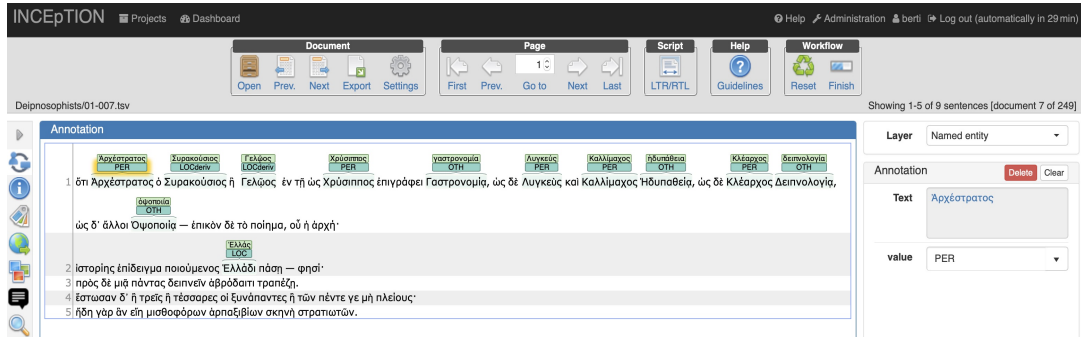


Figure 1: INCEpTION. Pre-annotated data (Ath., *Deipn.* 1.7).

A layer called *Ancient Greek Catalog* has been created in INCEpTION to annotate NEs that correspond to names of ancient authors and to descriptions/titles of ancient works, in order to disambiguate these entities and produce a text-based catalog of Greek literature with annotations of ancient Greek inflected forms and their corresponding lemmata. Figure 2 shows the *Ancient Greek Catalog* layer, whose values correspond to CTS URNs that uniquely identify authors and works (Berti et al., 2016). Individual entities have been linked together in spans corresponding to real entities, as for example Ἀρχέστρατος ὁ Συρακούσιος ἢ Γελῶσιος (Archestratus from Syracuse or Gela), who is identified as `urn:cts:greekLit:tlg1175`. In the same line, different forms of the title of Archestratus' work are identified with a CTS URN that also includes a reference to the author: `urn:cts:greekLit:tlg1175.tlg001`.¹⁶

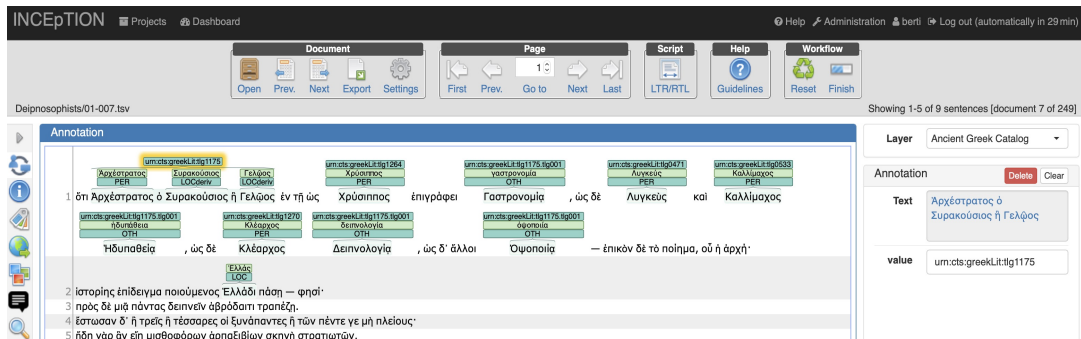


Figure 2: INCEpTION. *Ancient Greek Catalog* layer (Ath., *Deipn.* 1.7).

¹⁴https://webanno.github.io/webanno/releases/3.5.5/docs/user-guide.html#sect_webannotsv

¹⁵Sentence splitting has been performed on the basis of punctuation marks in ancient Greek: full stop (.), upper stop (·), and question mark (?).

¹⁶In these URNs, identifiers are based on the Canon of the Thesaurus Linguae Graecae (TLG): `tlg1175` is Archestratus and `tlg1175.001` is Archestratus' *fragmenta* (<http://stephanus.tlg.uci.edu/canon.php>).

5 Future Work

Future work will include complete NEs disambiguation and linking, and coreference resolution with a focus on entities related to ancient Greek authors and works. The final goal is to use INCEpTION for linking entity mentions to knowledge bases and structured vocabularies for ancient Greek authors and works that will enable scholars to annotate other texts and generate a text-based catalog of ancient Greek literature.

Acknowledgements

This work has been possible thanks to the support of the Alexander von Humboldt Stiftung (Digital Humanities Chair, Leipzig) and to a collaboration with the INCEpTION project.

References

- Alison Babeu. 2019. The Perseus Catalog: Of FRBR, Finding Aids, Linked Data, and Open Greek and Latin. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, pages 53–72. De Gruyter Saur, Berlin and Boston.
- Monica Berti, Christopher W. Blackwell, Mary Daniels, Samantha Strickland, and Kimbell Vincent-Dobbins. 2016. Documenting Homeric Text-Reuse in the *Deipnosophistae* of Athenaeus of Naucratis. *Bulletin of the Institute of Classical Studies*, 52(2):121–139.
- David Braund and John Wilkins, editors. 2000. *Athenaeus and His World. Reading Greek Culture in the Roman Empire*. University of Exeter Press, Exeter.
- Yanne Broux and Mark Depauw. 2015. Developing Onomastic Gazetteers and Prosopographies for the Ancient World through Named Entity Recognition and Graph Visualization: Some Examples from Trismegistos People. In Luca Maria Aiello and Daniel McFarland, editors, *Social Informatics. SocInfo 2014*, number 8852 in Lecture Notes in Computer Science, pages 304–313. Springer, Cham.
- Patrick Burns. 2019. Building a Text Analysis Pipeline for Classical Languages. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, pages 159–176. De Gruyter Saur, Berlin and Boston.
- Giuseppe G. A. Celano. 2019. The Dependency Treebanks for Ancient Greek and Latin. In Monica Berti, editor, *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, pages 279–297. De Gruyter Saur, Berlin and Boston.
- Richard Eckart de Castilho, Éva Mújdricza-Maydt, Seid Muhie Yimam, Sylvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann. 2016. A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84. The COLING 2016 Organizing Committee, Osaka, Japan.
- Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2017. Challenges and Solutions for Latin Named Entity Recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93. The COLING 2016 Organizing Committee, Osaka, Japan.
- Jan-Christoph Klie, Michael Bugert, Beto Bouldosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics, Santa Fe, New Mexico, USA.
- Damien Nouvel, Maud Ehrmann, and Sophie Rosset. 2016. *Named Entities for Computational Linguistics*. Focus Series. Wiley, London and Hoboken, NJ.
- Seth van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. 2015. Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections. *Digital Scholarship in the Humanities*, 30(2):262–279.

Enriching Lexicographical Data for Lesser Resourced Languages: A Use Case

Dirk Goldhahn

Natural Language Processing Group
University of Leipzig, Germany
dgoldhahn
@informatik.uni-leipzig.de

Thomas Eckart

Natural Language Processing Group
University of Leipzig, Germany
teckart
@informatik.uni-leipzig.de

Sonja Bosch

Department of African Languages
University of South Africa, South Africa
boschse@unisa.ac.za

Abstract

This paper presents a use case for enriching lexicographical data for lesser-resourced languages employing the CLARIN infrastructure. Basis of the presented work are newly prepared lexicographical data sets for under-resourced Bantu languages spoken in southern regions of the African continent. These datasets have been made digitally available using well established standards of the Linguistic Linked Open Data (LLOD) community. To overcome the insufficient amount of freely available reference material, a crowdsourcing Web portal for collecting textual data for lesser-resourced languages has been created and incorporated into the CLARIN infrastructure. Using this portal, the number of available text resources for the respective languages was significantly increased in a community effort. The collected content is used to enrich lexicographical data with real-world samples to increase the usability of the entire resource.

1 Introduction

The availability of contemporary text material is a prerequisite for a variety of applications and research scenarios, especially including studying recent developments in language use. Projects such as An Crúbadán¹ offer text freely available on the web to enable such studies for languages with small numbers of speakers. Resources in An Crúbadán are typically added manually, have a limited scope but are available for over 2,000 languages.

Crawling and processing Web content is now a standard procedure to acquire those needed resources. As a positive consequence of the vast amount of available online content, preselection of highly specific material is now possible for many languages and allows examination of all sorts of linguistic phenomena for specific domains and genres. Automatically generating valuable data sources out of online resources requires specific means of text acquisition and pre-processing of the gathered material. As a consequence, different systems that simplify the crawling and processing of Web pages for end users were developed and are in active use. One of the most popular services is the SketchEngine (Kilgarriff et al., 2014) which has a focus on lexicography.

On the other hand, for many lesser-resourced languages significant amounts of material are now available for the first time. Using standards of the growing Linguistic Linked Open Data (LLOD) community, connecting – so far isolated – datasets of all kinds helps creating substantial resources for these languages in a federated infrastructure. One of the benefits of these interconnections is the ability of building bridges between lexicographical entries and concrete, real-world examples of their use. The resulting resources

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

¹<http://crubadan.org/>

have a high value for all kinds of use cases and user groups, including being essential for first- and second-language acquisition.

This paper focuses on a specific use case, facilitated by the CLARIN infrastructure, in which newly created lexicographical datasets for some Bantu languages were enriched using a new Web crawling portal focusing on the acquisition of text material for lesser-resourced languages.

2 Crawling Under-Resourced Languages

Only for a small number of languages the situation concerning the availability of digital language resources is satisfactory. Even for most of the languages with more than one million speakers no reasonably sized textual resources or tools like POS taggers adapted to these languages are available. This points to a widespread need for digital language resources for many languages of the world. Therefore, a Web portal² for corpus collection with a focus on under-resourced languages with more than one million speakers has been initiated (Goldhahn et al., 2016) as a service in the CLARIN infrastructure. It relies on native speakers with knowledge of Web pages in their respective language. The initiative gives interested scholars and language enthusiasts the opportunity to contribute to corpus creation or extension by simply entering a URL into a Web interface.

In the backend, Heritrix (Mohr et al., 2004) the crawler of the Internet Archive is used in combination with a well established corpus processing chain that was adapted to append newly added Web pages to continuously growing corpora. This enables us to collect larger corpora for under-resourced languages by a community effort. These corpora are made publicly available within the CLARIN infrastructure, on the one hand directly in the portal and on the other hand as part of the Leipzig Corpora Collection.

Since its establishment, the Web portal helped creating initial text resources for several languages as well as expanding available text collections. All in all, more than 10,000 URLs were submitted to the system in 127 crawling jobs for 62 languages. Concrete efforts for two languages - Xhosa (ISO 639-3: xho) and Kalanga (ISO 639-3: kck) - will be described in the following sections.

3 Dictionary Data

Like most Bantu languages, Xhosa and Kalanga are considered resource scarce languages, implying that linguistic resources such as large annotated corpora and machine-readable lexicons are not available. Moreover, academic and commercial interest in developing such resources is limited. In the following section, available sources for lexicographical data for Bantu languages used in this publication are described in more detail.

3.1 Xhosa Dictionary

Xhosa lexical data was taken from a resource compiled by J.A. Louw (University of South Africa UNISA) which is available under a Creative Commons (CC) license. This Xhosa lexicographical data set consists of morphological information accompanied by English translations. It was created and made available by the authors for purposes of further developing Xhosa language resources (Bosch et al., 2018). The data were compiled with the intention of documenting Xhosa words and expanding existing bilingual Xhosa dictionaries by means of – among others – botanical names, animal names, grammar terms, modern forms etc., as well as lexicalisations of verbs with extensions. The publication process involved digitisation into CSV tables and several iterations of quality control in order to make the data reusable and shareable. Its current state is already published in the CLARIN infrastructure³ and available via a dedicated Web portal⁴.

3.2 Kalanga Dictionary

Lexicographical data for the Kalanga language was extracted from the Comparative Bantu OnLine Dictionary (CBOLD⁵). The project started in 1994 to create a source for lexicographical data for Bantu

²<https://curl.corpora.uni-leipzig.de/>

³<https://hdl.handle.net/11022/0000-0007-C655-A>

⁴<https://rdf.corpora.uni-leipzig.de>

⁵<http://www.cbold.ish-lyon.cnrs.fr/>

languages. The amount and range of available data, and its quality and format vary from dictionary to dictionary. The CBOLD dictionary for Kalanga was created in 1994 by Joyce Mathangwane and is provided as a plain text file. The dictionary contains 2960 lexemes with information about the part of speech, tone, noun classes and prefix/stem structure for the nouns. Additionally, English translations are provided.

3.3 Bantu Language Model

The lexical resources introduced were transformed into a unified schema to simplify all relevant data enrichment and quality assurance procedures and to form a basis for future applications and user interfaces. The Bantu Language Model (BLM) (Bosch et al., 2018) is an ontology of the Linguistic Linked Open Data (LLOD) community that ensures semantic and structural interoperability. The BLM is based on the MMoOn ontology (Klimek, 2017) and allows for the representation and interrelation of lexical, morphological and translational elements but also common grammatical meanings as well as noun class elements of Bantu languages.

The benefit of using an LOD-based format is the simplicity to enrich existing datasets with additional information. In the context of this contribution, the focus lies on connecting lexicographical data with real-world samples. The resulting interconnected resources are helpful for a variety of user groups and topics, including support of first and second language acquisition, or as a general resource for all types of text-producing activities.

4 Collecting and Processing Language Data

In a next step, the newly created lexical data were enriched with additional information: the availability of sample sentences proves to be valuable for users of lexical resources since they provide real-world usage examples of the lexical units. In the beginning of the project, available text data for the respective languages was very limited, both in the Leipzig Corpora Collection (LCC) and in other freely available digital resources. For Xhosa less than 18,000 sentences could be found in the LCC. For Kalanga the situation was even worse with only about 600 sentences. By advertising the initiative to some researchers in the respective communities, we were able to collect 180 seed URLs for Xhosa and one Web domain for Kalanga. Crawling resulted in 45,585 additional unique sentences for Xhosa and 996 for Kalanga, increasing available resources significantly. Figure 1 depicts these text collecting efforts for Xhosa.

The textual resources were processed to serve as a basis for assigning sample sentences to dictionary entries. For Xhosa this resulted in sample sentences for about 25% of the lexical entries available. Since the coverage is significantly higher for more frequent words and these words are typically queried more often, this will result on a higher sample sentence coverage for actual queries.

The textual data are made available in the CLARIN infrastructure via the Leipzig repository. This allows for download of the data sets, for searching in the data via FCS or the local Leipzig Corpora Collection portal, for sustainably citing textual resources on sentence level and for further processing using Web tools such as WebLicht. In a next step, extending the Bantu Language Model dataset is planned to allow for a direct linking of lexical entries and sample sentences using LLOD formats and therefore for an easier integration.

5 Conclusion and Further Work

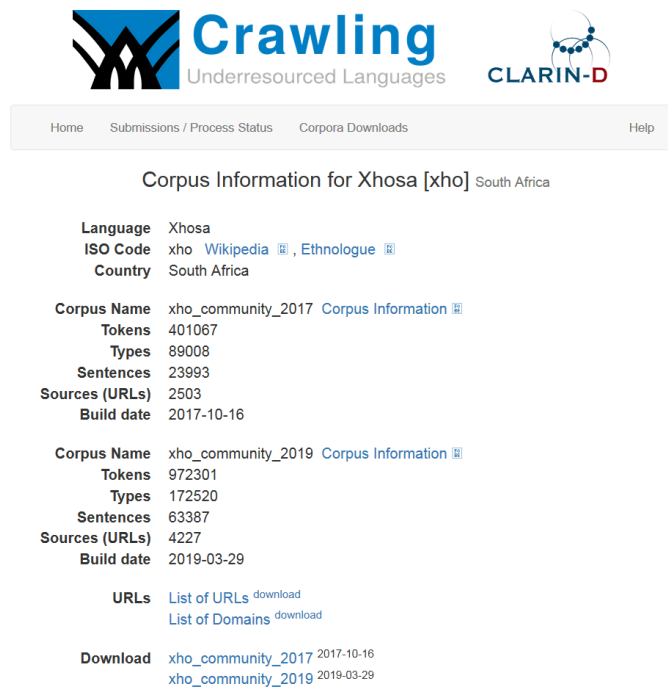
This paper presented a use case for enriching lexicographical data for lesser resourced languages with sample sentences. Basis were recently added resources and services of the CLARIN infrastructure such as Xhosa lexicographical data based on the Bantu Language Model⁶ and a portal for crawling lesser resourced languages (CURL⁷). Results are made available via the CLARIN infrastructure to allow for wide applicability. They include text corpora for Xhosa⁸ and Kalanga⁹.

⁶<https://hdl.handle.net/11022/0000-0007-C655-A>

⁷<https://hdl.handle.net/11022/0000-0007-D369-5>

⁸<https://hdl.handle.net/11022/0000-0007-D396-1>

⁹<https://hdl.handle.net/11022/0000-0007-D395-2>



Home Submissions / Process Status Corpora Downloads Help

Corpus Information for Xhosa [xho] South Africa

Language Xhosa
 ISO Code xho [Wikipedia](#) [Ethnologue](#)
 Country South Africa

Corpus Name xho_community_2017 [Corpus Information](#)
 Tokens 401067
 Types 89008
 Sentences 23993
 Sources (URLs) 2503
 Build date 2017-10-16

Corpus Name xho_community_2019 [Corpus Information](#)
 Tokens 972301
 Types 172520
 Sentences 63387
 Sources (URLs) 4227
 Build date 2019-03-29

URLs [List of URLs download](#)
[List of Domains download](#)

Download [xho_community_2017](#) 2017-10-16
[xho_community_2019](#) 2019-03-29

Figure 1: Overview of crawling activities in the CURL portal for Xhosa since 2017.

Future work will focus on deeper integration of the CURL portal into the CLARIN infrastructure. Advanced options for format conversion (e.g. TCF or plain text) are currently implemented. This will allow for direct processing of crawling results in environments such as WebLicht by employing the Language Resource Switchboard. Support for CLARIN's private work space solution will increase usability even further. In addition, the BLM-based RDF datasets will be extended to allow for direct integration of sample sentences and hence enhanced usage scenarios.

References

- Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn, and Uwe Quasthoff 2018. *Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment*, Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki (Japan) 2018.
- Thomas Eckart, Sonja Bosch, Dirk Goldhahn, Uwe Quasthoff, and Bettina Klimek 2019. *Translation-based Dictionary Alignment for Under-resourced Bantu Languages*, OpenAccess Series in Informatics (OASIS), Vol. 70: Language Data and Knowledge LDK 2019.
- Dirk Goldhahn, Maciej Sumalvico, and Uwe Quasthoff 2016. *Corpus Collection for Under-Resourced Languages with more than One Million Speakers*, CCURL 2016 Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity 2016.
- Adam Kilgarriff, Vit Baisa, Jan Buta, Milos Jakubicek, Vojtech Kova, Jan Michelfeit, Pavel Rychly, and Vit Suchomel 2014. *The Sketch Engine: ten years on*, Lexicography, pp. 7–36, Springer, 2014.
- Bettina Klimek 2017. *Proposing an OntoLex-MMoOn Alignment: Towards an Interconnection of two Linguistic Domain Models*, Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets. 2017.
- Gordon Mohr, Michele Kimpton, Michael Stack, and Igor Ranitovic 2004. *Introduction to heritrix, an archival quality web crawler*, Proceedings of the 4th International Web Archiving Workshop IAWW'04. 2004.

CLARIN-Supported Research on Modification Potential in Dutch First Language Acquisition

Jan Odijk

Utrecht University, the Netherlands

j.odijk@uu.nl

Abstract

This paper analyses data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning ‘very’, which show syntactic differences in modification potential. It continues the research reported on in (Odijk, 2016). The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular the treebank search applications *PaQu* (Parse and Query) and *GrETEL* Version 4.00. The analysis benefits from the use of parsed corpora (treebanks) in combination with the search and analysis options offered by *PaQu* and *GrETEL*. Earlier work showed that despite little data for *zeer* modifying adpositional phrases adult speakers end up with a generalised modification potential for this word. In this paper, we extend the dataset considered, and find more (but still little) data for this phenomenon. However, we also find a similar amount of data that form counterexamples to the non-generalisation of the modification potential of *heel*. We argue that the examples with *heel* concern constructions with idiosyncratic semantics and therefore are not counted as evidence for the general rule of modification. We suggest a simple statistical analysis to account for the fact that children ‘learn’ that *heel* cannot modify verbs or adpositions though there is no explicit evidence for this and they are not explicitly taught so.

1 Introduction

In this paper we analyse data to address a specific linguistic problem, i.e. the acquisition of the modification potential of the three more or less synonymous Dutch degree modifiers *heel*, *erg* and *zeer*, all meaning ‘very’ (Odijk, 2016). The analysis makes crucial use of linguistic applications developed in the CLARIN infrastructure, in particular the treebank search applications *PaQu* (Parse and Query, (Odijk et al., 2017)) and *GrETEL* Version 4.00 (Odijk et al., 2018), both of which make use of the Dutch syntactic parser *Alpino* (Bouma et al., 2001). The words that are being investigated are highly ambiguous. Most of the ambiguity is resolved by considering the syntactic context they occur in. Therefore the analysis benefits from the use of parsed corpora (treebanks). The data analysis process is considerably speeded up and facilitated by these parses in combination with the search and analysis options offered by *PaQu* and *GrETEL*.

2 The problem

The three Dutch words *heel*, *erg* and *zeer* are (near-)synonyms meaning ‘very’, i.e. (stated informally) they modify a word or phrase that expresses a (gradable) property or state and specify that its modifiee has the property or state it expresses to a high degree. Of these, *heel* can modify adjectival (A) phrases only, while *erg* and *zeer* can modify not only adjectival, but also verbal (V) and adpositional (P) phrases. For further details on the facts we refer to (Odijk, 2015; Odijk, 2016).

We take it for granted here that children ‘know’ or find out that the syntactic modification potential must be stated in terms of selection of a grammatical category (Odijk, 2019).

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

The first research question of this paper is: how can children acquire the differences in modification potential between these words? And, as we will see below, children will have to be able to generalise beyond the input sample on the basis of only a few occurrences of *zeer* modifying PPs, while they should not generalise for the same amount (or even more) occurrences of *heel* modifying PPs. The second research question is then: how do children make this distinction?

3 The Treebank search applications PaQu and GrETEL 4.0

It is important to investigate the use of these words in their syntactic context, because they are (as many words in natural language) highly ambiguous, as shown by Odijk (2016). The treebank search applications *Parse and Query (PaQu)* (Odijk et al., 2017) and *GrETEL* Version 4.00 (Odijk et al., 2018) make this possible.

4 Earlier work

The type of problem dealt with here has, at least for English phenomena, figured prominently in the language acquisition literature (Baker, 1979; Berwick, 1985; Pinker, 1989; Xu and Tenenbaum, 2007; Yang, 2016). We will discuss this work in the full paper.

Odijk (2015) analyses the Dutch CHILDES corpora (MacWhinney, 2000) for the words *heel*, *erg* and *zeer*.¹ The most important conclusions of this work can be summarised as follows: (1) there is an overwhelming number of cases where *heel* modifies an adjectival phrase (>92%); (2) modification of verbal or adpositional phrases by *heel* does not occur. (3) there are no clear examples with *erg* or *zeer* modifying a PP.

Odijk (2016) observes that *heel* occurs very early in the children's speech (1;11)², with *erg* occurring only a year later (2;10), and *zeer* very late (4;8). He ascribes the late occurrence of *zeer* to its more formal character. Odijk (2016) also investigates some adult corpora for these phenomena, in particular the written Dutch corpora LASSY (van Noord et al., 2013) and SoNaR (Oostdijk et al., 2013) and the Spoken Dutch Corpus (Oostdijk et al., 2002). His findings are as follows: (1) the LASSY data fully confirm the assumptions on the facts made in section 2; (2) the Spoken Dutch Corpus data contain examples in which *heel* modifies PPs; (3) the SoNaR data offer several examples where *heel* modifies PPs. (e.g. *heel in de verte*, lit. very in the distance, 'at a very great distance').

Summarising: (1) the data seem appropriate for acquiring the property that *heel* modifies adjectival but no verbal phrases; (2) it is less clear how modification of PPs can be excluded, since there are some examples where *heel* modifies PPs; (3) the absence of data for *zeer* makes it difficult to state anything about the acquisition of its modification potential.

In order to address the latter two problems, more data are needed. Unfortunately, there are no other CHILDES data for Dutch that are relevant in this context. Fortunately, there is another corpus that is relevant. The BasiLex corpus (Tellings et al., 2014) contains texts that are directed at children at primary school. The BasiLex corpus is significantly larger than the CHILDES corpora (11.5 million tokens). Because of the late acquisition of *zeer*, BasiLex's focus on texts that are targeted at children between the ages of 6 and 12 appears to make it particularly appropriate for investigating the modification potential of *zeer*.

5 Extending the dataset: the Basilex corpus

We used PaQu to investigate the properties of modifiees of *heel*, *erg* and *zeer*, respectively in Basilex.

Details of the query results and the manual analysis will be provided in the full paper. The results are summarised in Table 1. Strikingly, examples with *heel* modifying a PP are more frequent than *zeer* modifying a PP, but this does not have the effect that the adult grammar allows modification of PPs by *heel* in general. Conversely, despite their low frequency even in this larger corpus, the adult grammar allows modification of PPs by *zeer* generally.

¹These corpora contain approximately 2.5 million tokens.

²Using CHILDES notation for children's age: y;m means 'y years and m months'

Basilex	<i>heel</i>	<i>erg</i>	<i>zeer</i>	Wikipedia	<i>heel</i>	<i>erg</i>	<i>zeer</i>
Mod A	172	350.4	26.7	Mod A	90.5	128.3	342
Mod V	0	74.7	1.7	Mod V	0	12.2	18
Mod P	1.7	3.5	0.3	Mod P	0.3	2.1	1.9

Table 1: Relative frequency (per million tokens) of *heel*, *erg* and *zeer* as modifiers of adjectival (Mod A), verbal (Mod V) or adpositional (Mod P) phrases in the Basilex corpus and the Wikipedia part of the LASSY-Large corpus.

So, despite a larger and more representative corpus, the same question still lies before us: (1) why does the presence of PPs modified by *heel* not lead to generalising the modification potential of *heel* to PPs?; (2) why is the modification potential of *zeer* generalised to PPs despite its very low frequency?

Again we might ask ourselves whether the Basilex corpus is big enough to get a representative overview. So we turn to even larger corpora. Though these corpora are not representative at all for language acquisition, they may give us insight into the degree of representativity of the CHILDES corpora and the BASILEX corpora for the problem at hand.

6 Results from even larger datasets: LASSY Large and SoNaR

We investigated the modifiers of *heel*, *erg* and *zeer* in the Wikipedia part of the LASSY-Large corpus. This is a large corpus (145 million tokens) containing a very formal text genre (encyclopedia), so we expect *zeer*, being formal in nature, to occur more frequently than in the corpora investigated so far. This is indeed the case, see Table 1.

It is clear that *zeer* occurs much more often than in the earlier corpora as a modifier of adjectival, verbal and adpositional phrases. But even here, in this large and very formal corpus, the frequency of *zeer* modifying a PP is extremely low (1.9 per million). Examples of *heel* modifying a PP are less frequent in this corpus, but we ascribe this to the rather formal nature of this corpus.

A more detailed analysis of the SoNaR data yields slightly different figures but still do not help us resolve the research questions. (Details will be provided in the full paper).

7 Analysis of the research questions

We repeat the second research question addressed in this paper: (1) why does the presence of PPs modified by *heel* not lead to generalising the modification potential of *heel* to PPs?; (2) why is the modification potential of *zeer* generalised to PPs despite its very low frequency?

We suggest the following analysis. An expression such as *in de verte* ‘at a great distance’ indicates a location. A location is not gradable, hence cannot be semantically modified by words such as *heel*, *erg*, or *zeer*. We will provide supporting evidence for this claim from a totally independent construction.

If that is the case, then what does *heel* modify semantically in an expression such *heel in de verte*? As its translation indicates, *heel* semantically modifies the adjective (*ver*) inside a derived noun (*ver-te*) contained in a noun phrase itself contained in PP. No productive rule of grammar allows such an associated meaning operation, so this construction must be stored by the language learner as an idiosyncratic mapping between a form and a meaning.

We will provide additional evidence for the non-productive and idiosyncratic nature of the constructions in which *heel* modifies a PP, in particular the nouns cannot be replaced by semantically similar nouns and the adpositions that can occur come from a very small set.

We thus conclude that these constructions are treated as idiosyncratic form-meaning mappings (multiword expressions), and therefore their existence is not taken as evidence for the productive option for *heel* to modify PPs.

This now makes it possible to account for the first research question by using a simple statistical acquisition procedure. We assume that each property and rule in the grammar has an activation score. Whenever evidence for a rule or property is obtained (because use is made of it to analyse an input sentence or to produce a sentence), this activation score is increased. Only when the activation score of a rule

or property is above a certain threshold (to be determined empirically), the rule or property becomes an active element of the grammar. There is also a decay function that lowers the activation score regularly. Such a system has the potential to account for (1) acquisition of linguistic properties and rules based on input; (2) ill-formedness of phenomena for which insufficient input has been obtained and which are not subject to a productive rule (indirect negative evidence); (3) robustness of the acquisition device against ill-formed input, misheard or wrongly analysed utterances ((Yang, 2016, 13) and references there); (4) indeterminacy and/or variability of well-formedness judgements on specific data; (5) language attrition; (6) arising confusion among linguists about data they are analysing; (7) changing well-formedness judgements for new uses of words. Considering the examples being analysed here, for *erg* and *zeer* there is sufficient evidence in the data for the full modification potential (can modify A, V and P), while for *heel* there is only sufficient evidence for modifying A but not enough for modifying V or P to get above the threshold.

8 Concluding Remarks and Future Work

We will provide our full conclusions in the full paper. Here we mention only one conclusion: This paper formulated two linguistic research questions in the area of language acquisition and provided an analysis for them. The analysis made crucial use of syntactic search applications developed in CLARIN. The use of these applications was especially beneficial because of the high ambiguity of the words being investigated. This is the case even though the automatically created parses contain errors and require manual verification.

References

- C.L. Baker. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4):533–581.
- Robert Berwick. 1985. *The Acquisition of Syntactic Knowledge*. MIT Press, Cambridge, MA.
- Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edition.
- Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 23, pages 281–297. Ubiquity, London, UK. DOI: <http://dx.doi.org/10.5334/bbi.23>. License: CC-BY 4.0.
- Jan Odijk, Martijn van der Klis, and Sheean Spoel. 2018. Extensions to the GrETEL treebank query application. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 46–55, Prague, Czech Republic, January 23-24. <http://aclweb.org/anthology/W/W17/W17-7608.pdf>.
- Jan Odijk. 2015. Linguistic research with PaQu. *Computational Linguistics in the Netherlands Journal*, 5:3–14, December.
- Jan Odijk. 2016. A Use case for Linguistic Research on Dutch with CLARIN. In Koenraad De Smedt, editor, *Selected Papers from the CLARIN Annual Conference 2015, October 14-16, 2015, Wroclaw, Poland*, number 123 in Linköping Electronic Conference Proceedings, pages 45–61, Linköping, Sweden. CLARIN, Linköping University Electronic Press. <http://www.ep.liu.se/ecp/article.asp?issue=123&article=004>, <http://dspace.library.uu.nl/handle/1874/339492>.
- Jan Odijk. 2019. The power of syntactic selection in language acquisition. unpublished paper, Utrecht University <https://surfdrive.surf.nl/files/index.php/s/ITRD6L4s5NuFeP3>, February.
- N. Oostdijk, W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen. 2002. Experiences from the Spoken Dutch Corpus project. In M. González Rodríguez and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 340–347. ELRA, Las Palmas.

- N. Oostdijk, M. Reynaert, V. Hoste, and I. Schuurman. 2013. The construction of a 500 million word reference corpus of contemporary written Dutch. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, pages 219–247. Springer, Berlin. <http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1>.
- Steven Pinker. 1989. *Learnability and Cognition*. MIT Press, Cambridge, MA.
- Agnes Tellings, Micha Hulsbosch, Anne Vermeer, and Antal van den Bosch. 2014. BasiLex: an 11.5 million words corpus of Dutch texts written for children. *Computational Linguistics in the Netherlands Journal*, 4:191–208, 12/2014.
- Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.
- F. Xu and J.B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review*, 114(2):245.
- Charles Yang. 2016. *The Price of Productivity: How Children Learn to Break the Rules of Language*. MIT Press, Cambridge, Mass.

Training Workshops in the Bi-directional Model of the Language Technology Infrastructure Development

Maciej Piasecki

Dep. of Computational Intelligence
Wrocław University of Science
and Technology, Poland
maciej.piasecki@pwr.edu.pl

Jan Wiczorek

Dep. of Computational Intelligence
Wrocław University of Science
and Technology, Poland
jan.wiczorek@pwr.edu.pl

Abstract

In this paper we describe the evolution of training workshops offered by the CLARIN-PL. We focus on the types of workshops, the competences of participants and the role which the workshops are aimed to fulfil in a bi-directional model of the language technology infrastructure development assumed for CLARIN-PL. The paper also discusses our experience collected during four years and examples of the influence of the workshops on users and their cooperation with CLARIN.

1 Users as a Basis for Language Technology Infrastructure

Language Technology Infrastructure (LTI) is a complex system that enables combining language tools with language resources into processing chains (or pipelines) with the help of software framework. The processing chains are applied to language data to obtain results interesting from the perspective of research needs of users. Addressing user needs is a fundamental challenge in developing computer systems. Users make all systems imperfect, but also users are a reason for a computer system to exist. Moreover, LTI must have users to be funded and LTI for Social Sciences & Humanities (SS&H) must attract users who are significantly different from its developers, i.e. language engineers.

Ideally, the user should be present and in focus during all stages of system development. In a system development process which is user-driven and holistically focused on user experience, the *context of use* (i.e. users, their tasks and the environment) determines the perspective from which the users perceive LTI, i.e. in the perspective of its *usability* (efficiency, effectiveness and satisfaction) (ISO 9241).

Different CLARIN members follow different schemes of LT development. However, three main approaches can be distinguished (Piasecki, 2014). The first, a *bottom-up* process concentrates on collecting already existing language tools and resources (LTRs) and is aimed at establishing accessibility and technical interoperability of LTRs. As a result, the tools and resources become accessible via Web to the users and can be combined into processing chains. The second, a *top-down* model is entirely focused on complete research applications (or tools) for the final users. It starts with collecting descriptions of users' tasks and requirements for research applications. Next the underlying network system of services and LT components should be designed and developed according to the requirements. It is potentially attractive, but highly unrealistic, at least because of high workload (and cost) required, while the outcome is uncertain and deferred in time.

The third one, a *bi-directional* process (Piasecki, 2014) starts with combining several existing LTRs into an infrastructure, too, but taking into account user-driven requirements for lacking LTRs. Moreover in this approach, designing top level research applications for users is a starting point for many activities in the LT infrastructure development. The bi-directional approach seems to be also natural due to the fact that SS&H research methods are still evolving or being shaped, and LT can inspire this process.

The bi-directional approach requires effective communication between LTI developers and users. In this paper we want to analyse a mechanism of training workshops as a means for this communication.

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

2 Ways to Users

There is a broad range of possible dissemination and communication methods, like advertisements, on-line and printed materials, publications, lectures etc. However they share two potential drawbacks: are focused on promoting a completed product and mainly work in one direction - from developers to potential users. These drawbacks are of special concern in the light of limited funding for the whole infrastructure. The bi-directionality requires feedback from direct collaboration and live contact with users, including *key users* among them.

Since 2015, CLARIN-PL has organised many training workshops addressed to different groups of researchers in SS&H. Motivations, topics and organisational forms of the workshops were diversified and to a very large extent determined by the needs of the cooperating partner institutions. Three types of workshops have emerged:

- *General Workshops* presenting an overview of the contemporary state of CLARIN-PL and aimed at broad scientific audience (e.g. subsequent editions of training workshops “CLARIN-PL in Research Practice”)
- *Targeted Workshops* – focused on selected subdomains, types of research task or even users’ subgroups or sub-communities, organised for much smaller number of participants,
- *User-defined Workshops* – invited by the users, sometimes originating from their very concrete needs, sometimes from curiosity.

3 General Workshops

General workshops are large events, typically between 60 to 120 participants, and are aimed at making SS&H researchers aware of the LT existence and its possible research applications, as well as the CLARIN-PL offer of services and tools. The main intended attendees – researchers and PhD students from different SS&H disciplines – can represent very diversified competence in LT. With such general audience we wanted not only to increase the awareness of possible LT applications, but also to broadly look for potential active users (researchers) of our CLARIN-PL LTI. That is why we started such workshops at an early stage of the development in 2015 having completed only a couple of prototypes.

Among the well-known barriers preventing the use of LT in SS&H, namely technical, knowledge and skill, legal and financial, the first one appeared to be most severe because of gaps observed in the robust LT for Polish. CLARIN-PL contributed a lot to overcoming this and other barriers. However, yet another negative effect could be observed and it still is the case. Low awareness of the Polish LT potential for SS&H causes a situation in which most Polish SS&H researchers are not able to formulate their research tasks in way taking into account the use of LT or even to describe their expectations towards potential results from LT application. This is in clear contrast to intensive contact with users which is required in the bi-directional development model to recognise their needs and make the users involved during the evaluation of the constructed tools and services. In order to improve this situation, in late 2014 (about one year since the CLARIN-PL construction phase had been initiated), we started regularly organising large training workshops. In workshop composition we tried to reserve only limited amount of time for lectures (no more than half) preserving an ample time for practical training classes with hands on our web applications (in fact prototypes during the first workshop editions). The first three editions of General Workshops were organised between April and June 2015 in three large cities and scientific centres: Warsaw, Cracow and Wroclaw.

A typical General Workshop lasts for 2-3 full days, often with two parallel streams of lectures and training classes organised in thematic blocks. Each workshop starts with an introduction to CLARIN (basic ideas, network of centres, VLO, Federated Content Search), Polish centre and repository. Next, we present an overview of LTRs, services and web applications offered by CLARIN-PL. We provide basic technical information about the infrastructure, its limitations and research projects in which it has already been used. However, most of the time is devoted to thematic blocks like: Polish corpora (construction, annotation and searching), bilingual and monolingual corpora, lexicographic resources and tools, extraction of multiword expressions and terminology, basic statistical analysis of texts, speech corpora and speech processing, stylometry, semantic analysis of texts or grammatical analysis (parsers and statistics). Each block is focused on concrete tools offered by CLARIN-PL.

Workshops of this type were organised nine times in the years 2015-2018, with a total of 700 participants (that is a relatively high number, when we take into account that the workshops are run in

Polish). We gradually started to reduce the frequency of such workshops down to 1-2 per year, because not all our expectation have been fulfilled. Positives, awareness about CLARIN-PL and Polish LI in SS&H research increased significantly. For instance, reports were published by workshop participants in a scientific journal (Redzimska J., Stanulewicz D., Wawrzyniak-Śliwska M., [report] *CLARIN-PL workshops and lectures in Gdańsk, 18–19 May 2018*²) or a peer-reviewed publication presenting an overview of NLP tools for H&SS (Beta Duda, Karolina Liszczyk, Digital Tools in the Polish Language Academic Didactics - Applications, Possibilities, Prospects³)

We established cooperation with several key users (who are also very supportive in the development of tools) and many contacts across different SS&H areas. As a snowball effect, new invitations to organise workshops are coming. Negatives, only a small number of users stay active after a workshop. The time during the workshop is too small for different issues and the topics are too diversified. The audience seems to be too diversified according to their background, skills and research tasks. The last factor forces us to invent some more or less artificial or too simple use examples and example data for applications. In addition, a large total number of participants and limited time result in groups of 20-30 participants during practical classes that are unmanageable with respect to the individual help (that is often very required, e.g. the lack of knowledge and skills in LT). The impact on the uptake of CLARIN-PL in research projects was also smaller than expected. Moreover, the increase of spontaneous users of web applications is not as large, as we could expect.

4 Targeted Workshops

After two years and several editions of General Workshops, we had already established contacts with research teams who were making first steps or were planning to use LT in their research. An idea of *Targeted Workshops* originated mainly from these contacts.

The main objective of a targeted workshops is to improve the participants' competence and to train them up to the level of working alone with CLARIN-PL resources and services. Researchers attending such a workshop should share similar disciplines, research objectives or types of language data. Workshop groups shall not exceed 24 persons, especially during training classes. Participants shall be experienced in their fields, but are not expected to have skills in LT. Ideally, an initiative to organise a workshop is based on cooperation between the users and CLARIN-PL. It can originate from prior support for the users' research tasks or contacts established during other workshops. Classes conducted during the targeted workshops are very focused thematically. They usually concern a coherent set of issues. We encourage active involvement of participants, who are asked to perform some concrete tasks with the help of CLARIN-PL tools. It is very good if the tasks (assignments) can be defined in relation to the materials and research questions provided by the inviting users.

Examples: a workshop on using the DSpace repository and information extraction tools (LEM, WebSty) for researchers analysing the statements of Polish MEPs (July 2018, Warsaw), a workshop on the use of the repository and corpus tools, or a workshop for psychologists (May 2019).

Targeted workshops result in significant increase in the participants' competences and users' activity in the centre, e.g. a discourse researcher team (Univ. of Silesia) participated in a general workshop (Feb. 2017, Łódź). Next, they invited CLARIN-PL to provide workshop on the use of language technology in diachronic linguistics during a conference on Polish diachronic corpora (Apr. 2017, Katowice). In autumn 2019, this research team is going to submit two applications for research grants. The preparation of the applications was preceded by a series of consultations and training provided by CLARIN-PL.

CLARIN-PL employees can also use the workshops to test new functions of the tools and collect new requirements. In the long term, the trained research teams may become independent and be able to formulate tasks that can be helped by LT.

² <http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-0460110e-4111-4f6b-b013-0f0cd18012b2>

³ <https://www.journals.us.edu.pl/index.php/FL/article/view/7432>

5 User-defined Workshops

Workshops whose content is determined by requirements formulated by the inviting partners are becoming more frequent with every year. They often result from targeted workshops. Participants are usually a narrow group of specialised researchers: no more than 20, sometimes much smaller group.

User-defined workshops resemble targeted workshops in many ways. However, the main difference is that they are initiated by external partners who have also significant influence on their programme. The program of such a workshop is focused on the given subject matter and the needs of the co-organiser. The main organisers of such events are usually scientific institutions of the users. In some cases, researchers initiating workshops contact CLARIN-PL on the basis of a recommendations of other CLARIN ERIC partners (SADiLaR) or Polish researchers already experienced in LT in SS&H (Polish-Yiddish dictionary, EngHum workshops in Warsaw - November 2018).

Practical classes usually concern a narrow topic and data defined by the main organizer. CLARIN-PL employees often cooperate with users in co-leading classes. The training itself is usually very proactive: research tasks are solved on the basis of material previously submitted by the organizer, with which the participants are also familiar. Examples: workshops conducted for the needs of SADiLaR on building the African Wordnet (Pretoria, February 2019), training on the use of the DSpace repository and the Kontext corpus search engine for employees of the Institute of World Art Studies (Warsaw, January 2019). The effects of such workshops are noticeable almost immediately, because a workshop is an element of research or preparations preceding the start of a users' project.

6 Workshop Organisation Procedure

A) Initiative and contact:

the organisation of the workshop is determined by the way in which they are initiated. In the case of *general workshops*, the initiative is always taken by CLARIN-PL, which defines the thematic scope of the programme and selects a convenient place (some region of Poland, gradually increasing coverage). Next, we are looking for a partner who has appropriate infrastructure (rooms, Internet access⁴, etc.) and is potentially interested in the subject of the workshops. The co-organizers usually come humanities, philology and social sciences faculties. The organisation of *targeted* and *user-defined workshops* usually results from the previous contacts, e.g. scientific events with presentation of research done with CLARIN-PL tools or events during which CLARIN-PL tools were presented or recommended, or authors mentioning CLARIN, or a quotation in an article or book.

B) Identifying the needs of participants:

the programme of the workshop, the type of activities, the way they are conducted is strictly determined by the needs and expectations of future participants.

For general workshops, the organisers assume low LT competence of the participants and also lack of their specific expectations. The organisation of other types of workshops requires an interview with experts (partners) who know the needs and competences of a specific (potential) participants. This interview takes the form of a dialogue in which partners from SS&H describe their needs and the CLARIN-PL staff select the appropriate tools/resources that may suit these criteria. After demonstrating the action model and discussing the possible role of LRSs and applications in the research process, a joint decision is made whether to include the topic in the workshop agenda.

C) Setting up the programme:

based on the identified and defined needs of the participants, the programme of a targeted and user-defined workshop is created. It is focused on one or two thematic threads (e.g., information extraction from texts and tools supporting lexicographic work or tools supporting translator's work and creation and annotation of language corpora), which can provided in parallel or separately. During the general workshop all activities are focused on demonstrating the tools' functionality, their possible applications and use examples.

⁴ This is a crucial issue: large number of users connecting several devices to the network in the same time and using communication intensive CLARIN-PL web applications has been a serious challenge for most institutions

References

- Henriksen, Lina; Hansen, Dorte Haltrup, Maegaard Bente, Pedersen Bolette S. and Povlsen Claus. 2014. Encompassing a Spectrum of LT Users in the CLARIN-DK Infrastructure. In ed. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. *Proceedings of the 9th International Conference on Language Resources and Evaluation : LREC 2014*, pp. 2175-2181.
- Offersgaard Lene, Jongejan Bart, Hansen Dore H. 2013. CLARIN-DK – Status and Challenges, *Proceedings of the Workshop on Nordic Language Research Infrastructure at NODALIDA 2013*. Linköpings universitet : Linköping University Electronic Press, pp. 21-32.
- Piasecki Maciej. 2014. User-driven Language Technology Infrastructure – the Case of CLARIN-PL. In ed. Tomaž Erjavec and Jerneja Žganec Gros. *Proceedings of the 9th Language Technologies Conference - Information Society - IS 2014, Ljubljana, Slovenia*, pp. 7-13.
- Redzimska Joanna, Stanulewicz Danuta, Wawrzyniak-Śliwska Magdalena. 2017. CLARIN-PL Workshops in Łódź, 3-4 February 2017. *Beyond Philology* 14(1): 251-257.
- Wynne Martin. 2013. The Role of CLARIN in Digital Transformations in the Humanities. *International Journal of Humanities and Arts Computing*, 7(1-2) 89-104.

OpeNER and PANACEA: Web Services for the CLARIN Research Infrastructure

Davide Albanesi

Istituto di Linguistica Computazionale
“A. Zampolli” (CNR-ILC)
Via G. Moruzzi, 1 - 56124 Pisa (ITALY)

Riccardo Del Gratta

Istituto di Linguistica Computazionale
“A. Zampolli” (CNR-ILC)
Via G. Moruzzi, 1 - 56124 Pisa (ITALY)

name.surname@ilc.cnr.it

Abstract

This paper describes the necessary steps for the integration of OpeNer and PANACEA Web Services within the CLARIN research infrastructure. The original Web Services are wrapped into a framework and re-implemented as REST APIs to be further exploited through both Language Resource Switchboard and WebLicht and made available for the CLARIN community.

1 Introduction and motivation

OpeNer and PANACEA¹ were two European projects funded within the 7th Framework Program and covering 4 years of research initiatives on Language Resources and Technologies (LRT). OpeNer developed some Natural Language Processing (NLP) tools in order “to detect [...] entity mentions [...]”, by “[...] performing sentiment analysis and opinion detection on” specific textual resources, especially in reviews for hotel accommodations and tourism at a large. Such tools were designed to be easily customizable for Academia, Research and Small and Medium Enterprises. The exhaustive list of services and lexicons developed by OpeNer as well as of the European languages covered are available at their official github.² The PANACEA project addressed “the most critical aspect of Machine Translation (MT): the so called language resource bottleneck.” PANACEA developed a set of linguistic resources, more precisely “a ‘factory’ of Language Resources (LRs) in the form of a production line [...]”, to automate the stages for “[...] acquisition, production, maintenance and updating of the language resources required by machine translation”. The platform created in the framework of PANACEA is a virtual and distributed environment where various interoperable components can be concatenated to create specific workflows to produce several language resources in various languages. The services developed in PANACEA are of great importance for Academia, Research and Small and Medium Enterprises, especially the ones focused on MT and related technologies. OpeNer and PANACEA share many aspects: from the creation of annotated corpora and lexicons to the development of web tools and services used to analyze and build them up to the focus on specific communities. They also share the concept of interoperability as the use of the Kyoto Annotation Format (KAF) (Bosma et al., 2009) in OpeNer,³ the Graph Annotation Format (GrAF) (Ide and Suderman, 2007) in PANACEA⁴ and the Lexical Markup Framework (LMF) (Francopoulo et al., 2006) in both of them states. Data and tools interoperability is a strategic goal in CLARIN.⁵ And, within CLARIN, initiatives such as the Language Resource Switchboard (LRS) (Zinn, 2018) and WebLicht (Hinrichs et al., 2010) openly go towards methodologies and “systems” addressing interoperability issues between language tools and language resources. These initiatives are central in CLARIN which therefore becomes the ideal environment for the tools and Web Services offered by OpeNer and PANACEA. Lastly, the services developed within both projects and the results achieved play

¹Respectively <http://www.opener-project.eu/> and <http://www.panacea-lr.eu/>

²<https://github.com/opener-project>

³<https://github.com/opener-project/kaf/wiki/KAF-structure-overview>

⁴http://www.panacea-lr.eu/system/graf/graf-T02_documentation_v1.pdf

⁵See, for instance

<https://www.clarin.eu/event/2019/parlaformat-workshop>, <https://www.clarin.eu/event/2017/clarin-workshop-towards-interoperability-lexico-semantic-resources-among-others>.

a key role for the CLARIN community as well. Indeed, on the one hand, the Virtual Language Observatory (VLO)⁶ contains several LRs but only some specific tools for Sentiment Analysis, while, on the other hand, many LRT are available for MT. This clearly means that the latter community is already in the CLARIN community, while the former one should be helped to get more involved. It is a matter of fact that the more the tools are published and used in and through CLARIN, the more they will have an impact on the community, and this community will tend to grow. However, it is obvious that the community involvement can not be managed only from the technological point of view, but a political point of view is needed as well.

2 Current Architecture and Common Aspects

ILC4CLARIN, hosted at the National Council of Research (CNR) “Institute for Computational Linguistics A. Zampolli” in Pisa, is the first and leading CLARIN B-centre of Italian Consortium, CLARIN-IT.⁷ ILC4CLARIN is already offering some of the PANACEA Web Services through a man-machine interaction available at <https://ilc4clarin.ilc.cnr.it/en/services/>, while OpeNER Web Services are offered through a local installation.

Tokenizer	http://opener.ilc4clarin.ilc.cnr.it/tokenizer	Pos Tagger	http://opener.ilc4clarin.ilc.cnr.it/pos-tagger
Kaf2Json	http://opener.ilc4clarin.ilc.cnr.it/kaf2json

Table 1: OpeNER Endpoints

In addition, the initial implementation of the tools has many common aspects, such as, for instance, the following: i) many tools in OpeNER and PANACEA are command line tool. To wrap command line tools, OpeNER uses Ruby⁸ and builds REST Web Services, while PANACEA uses Soaplab⁹ and offers SOAP Web Services; ii) OpeNER offers both POST and GET API; iii) PANACEA offers SOAP Web Services through a Web Interface; iv) simple pipelines are available in OpeNER, while a workflow engine¹⁰ is used in PANACEA; v) Kyoto Annotation Format (KAF), Lexical Markup Framework (LMF) and Graph Annotation Format (GrAF) guarantee the interoperability among data and services at different levels. Although their architectures differ, both foster interoperability as it is shown in the Figures 1 and 2.

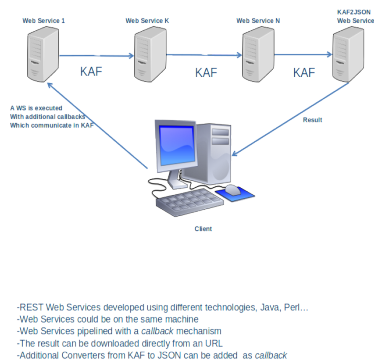


Figure 1: OpeNER Architecture

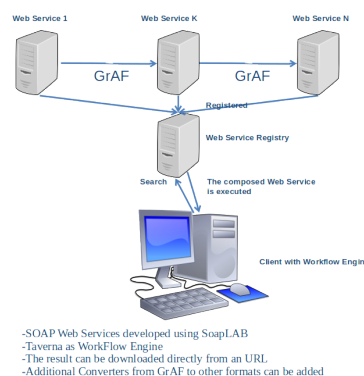


Figure 2: PANACEA Architecture

The images reported above show that the tools are ready to be inserted into workflows but, when it comes

⁶<https://vlo.clarin.eu/>
This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

⁷In order <https://ilc4clarin.ilc.cnr.it/>, <http://www.ilc.cnr.it/en/>, <https://www.clarin-it.it>

⁸See <https://rubyonrails.org/>

⁹Soaplab is described at <http://soaplab.sourceforge.net/soaplab2/>

¹⁰See Taverna, www.taverna.org.uk

to fully satisfy the requirements of Language Resource Switchboard (LRS) and WebLicht, we need to create new wrappers around the available tools so that we can correctly manage REST APIs that are able to “consume” and/or produce different formats. LRS does not require that the handled tools have specific output formats but, in any case, it requires the tools are able to read texts from URLs, uploaded files, or simple input boxes, while WebLicht accepts tools able to read and write the TCF format.¹¹

3 Moving OpeNer and PANACEA into CLARIN

In this section we describe the strategy used to integrate OpeNer and PANACEA Web Services into the CLARIN infrastructure. This initiative is carried out by the development team at the ILC4CLARIN centre.

3.1 Technical Implementation

Firstly, we have to consider that OpeNer offers REST Web Services (by default managed by APIs), while PANACEA offers SOAP ones. This means that the former is easily wrapped into a REST context, while the latter needs to be managed with some SOAP APIs before being inserted in a REST context. We have two alternatives: either i) to use the SOAP APIs provided by SoapLab;¹² or ii) to start from the original command line programs and use a different framework to transform these scripts into REST APIs. Both alternatives have their pros and cons. The first one forces the development team to code a shell around SOAP Web Services so that they can be “executed” by a software program and not just by a web interface (as it is the current case). Fortunately, the SoapLab APIs do exactly this: they provide methods to access Web Services in a simple way, without the burden of a SOAP implementation. Anyway, they are an additional piece of software to manage: the compatibility with existing libraries, methods that become deprecated and must be replaced, security flaws ... must be addressed. The second option needed to replicate (in a sense) what was already obtained through SoapLab. For instance, a framework like CLAM¹³ allows developers to transform command line programs into REST Web Services; and this would align the PANACEA and OpeNer Web Services. However, there would be a duplication in terms of both service endpoints and machines for hosting them. We simply considered this alternative non-economic (at least from our point of view), so we opted for alternative i). There are several strategies for implementing REST services; the majority follows the JAX-RS¹⁴ specifications. We decided to use the DropWizard¹⁵ framework which was described to the second author directly by the WebLicht developers during a hackaton¹⁶ held in Ljubljana in 2016. This framework combines an HTTP server (Jetty), a library for JAX-RS (Jersey), a “lingua franca” (Json¹⁷) and many other useful development tools. Furthermore, DropWizard is very useful for decoupling the basic implementation (the actual piece of code performing the operations) from how the code and the results of the operations are managed through the HTTP protocol. It is easy to understand that the situation of the PANACEA and OpeNer Web Services is exactly the following one: a core part (SOAP and REST Services respectively) and a wrapper that makes them accessible to programs through the HTTP protocol. In conclusion, DropWizard is a framework which helps to decouple the core tools from the corresponding web resources, as it is described in Listing 1.

¹¹The TCF format is described at https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format

¹²Java implementation in *taverna-soaplab-client* from the maven repository <http://www.mygrid.org.uk/maven/repository/>

¹³<https://proycon.github.io/clam/>

¹⁴<https://jcp.org/en/jsr/detail?id=339>

¹⁵<http://www.dropwizard.io/1.3.4/docs/>

¹⁶<http://www.clarin.si/info/events/mcat-workshop/>

¹⁷In order <https://www.eclipse.org/jetty/>, <https://jersey.github.io>, <https://www.json.org/>

```
/**
 * This is an example code which uses dropwizard, jetty and weblicht api to include
 * a core tool as a web service resource.
 ** * a lot of other imports */
import io.dropwizard.Configuration;
/**
 * the resource for TCF is something as
 */
@Consumes(TEXT_TCF_XML)
@Produces(TEXT_TCF_XML)
public StreamingOutput myexample(final InputStream text) {
    OutputStream tempOutputData = null;
    .....
    /* call the core */
    calltheservice(lang, text, some_other_parameters, tempOutputData);
    .....
}
/** the core where the actual tool is executed */
public void calltheservice(String lang, Map map, InputStream i, OutputStream o) {
    MyTool tool = new MyTool();
    TextCorpusStreamed textCorpus = null;
    .....
    tool.doSomething(textCorpus);
    ...
}
}
```

Listing 1: DropWizard integration

In Listing 2, a Jersey client executes the OpeNer web Service at one of its endpoint (endpoints are listed in Table 1) and returns the response.

```
/* import jersey stuff */
public doSomething(...) {
    client = Client.create();
    webResource = client.resource(URL_ENDPOINT);
    response = webResource.type(MediaType.APPLICATION_FORM_URLENCODED)
        .post(ClientResponse.class, formData);
    output = response.getEntity(String.class);
    setOutputStream(output);
    .....
}
}
```

Listing 2: Opener Snippet

```
/* import soaplab api stuff */
public doSomething(...) {
    SoaplabBaseClient client = getClient(SERVICE_ENDPOINT);
    SoaplabMap results = client
        .runAndWaitFor(SoaplabMap.fromMap(getInputs()));
    .....
    Map outputs = SoaplabMap.toMap(results);
    /* manage outputs and format */
    .....
}
}
```

Listing 3: Panacea Snippet

Listing 3 is quite similar to 2 but, in the latter, a client based on SoapLab APIs is responsible for executing the SOAP Web Service at the SERVICE_ENDPOINT. Soaplab APIs return the response that contains the analyzed text. The full code is available at <https://github.com/cnr-ilc/linguistic-tools-for-weblicht/tree/master/OpeNerServices> and <https://github.com/cnr-ilc/linguistic-tools-for-weblicht/tree/master/PanaceaServices>

However, in both examples, the response contains the analyzed text in the *tool native format*, which is KAF for OpeNer tools but can be any format for PANACEA ones. This is a PANACEA specific feature: actually, SoapLab simply wraps the original tools and produces the native output format, requiring only the implementation of converters to switch from one format to another.¹⁸ Our implementation is

¹⁸For example, the Freeling service returns a tabbed file which must be converted to other formats such as TCF, KAF ...

not limited to simply wrap the offered Web Services; additional endpoints have been added to manage different input and output formats. This is required to fully integrate our tools in both Language Resource Switchboard and WebLicht. Since input and output formats can be plain texts, TCF and KAF documents, the following POST and GET¹⁹ Web Services “consume” plain, TCF and KAF documents to produce TCF, TAB (tabbed) and KAF valid documents:

POST Plain texts `openerservice/tokenizer/runservice, panaceaservice/freeling_it/runservice`

POST TCF and KAF documents `openerservice/tokenizer/[tcflkaf]/runservice, panaceaservice/freeling_it/[tcflkaf]/runservice`

GET Plain texts `openerservice/tokenizer/lrs, panaceaservice/freeling_it/lrs`

GET TCF and KAF documents `openerservice/tokenizer/[tcflkaf]/lrs, panaceaservice/freeling_it/[tcflkaf]/lrs`

4 Conclusion and Future Work

In this paper we described an initiative carried out at ILC4CLARIN, which aims at integrating Web Services from PANACEA and OpeNer into the CLARIN infrastructure. The overall strategy is to wrap existing Web Services within REST APIs, so that both Language Resource Switchboard and WebLicht can exploit the new services. We have included only *Freeling* for Italian from the PANACEA set of Web Services and a basic tokenizer from OpeNer one. However, this work helped us to structure the various building blocks (see the github repository) so that other services can be easily wrapped, being them from PANACEA or OpeNer.

Truth be told, ILC4CLARIN already offers a tokenizer²⁰, which is a Java porting of the original OpeNer tokenizer. This service is in WebLicht²¹ and was successfully tested for Language Resource Switchboard. So, what are the reasons that led us to use a different strategy? For PANACEA nothing different had to be done while, for OpeNer, the use of the java porting technique required that every original service was rewritten in Java, regardless of the original programming language. This was essentially the motivation that led us to decide to wrap the original services instead of rewriting them.

ILC4CLARIN uses dockers²² for the services it offers and publishes the images in dockerhub²³. We will follow this trend for the new services as well.

References

- Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. 2009. KAF: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*.
- Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. 2006. Lexical Markup Framework (LMF) for NLP Multilingual Resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, MLRI '06, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hinrichs, M., Zastrow, T., and Hinrichs, E. 2010. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- Ide, N. and Suderman, K. 2007. GrAF: A Graph-based Format for Linguistic Annotations. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Zinn, C. 2018. The language resource switchboard. *Comput. Linguist.*, 44(4):631–639, December.

¹⁹GET endpoints have been set up for eventual integration into Language Resource Switchboard (LRS)

²⁰<https://ilc4clarin.ilc.cnr.it/services/ltfw-it/>

²¹<http://hdl.handle.net/20.500.11752/ILC-85@format=cmdi>

²²We use rancher (version 1 (<https://rancher.com/>)) to manage docker images and composition

²³<https://hub.docker.com/r/cnrilc/ltfw>

CLARIAH Chaining Search: A Platform for Combined Exploitation of Multiple Linguistic Resources

Peter Dekker

Dutch Language Institute
Leiden, The Netherlands
peter.dekker@ivdnt.org

Mathieu Fannee

Dutch Language Institute
Leiden, The Netherlands
mathieu.fannee@ivdnt.org

Jesse de Does

Dutch Language Institute
Leiden, The Netherlands
jesse.dedoes@ivdnt.org

Abstract

In this paper, we introduce *CLARIAH chaining search*, a Python library and Jupyter web interface to easily combine exploration of linguistic resources published in the CLARIN/CLARIAH infrastructure, such as corpora, lexica and treebanks. We describe the architecture of our framework and give a number of code examples. Finally, we present a case study to show how the platform can be used in linguistic research.

1 Introduction

Linguistic resources, such as lexica, corpora and treebanks, are usually published as web applications, where users issue a search term and a number of results are shown in the browser. However, connecting multiple web applications in a single process for research and analysis is a difficult task. Of course, some linguistic resources provide raw access to the data in an export file or through an API. But this mostly requires in-depth knowledge of software development and is labour-intensive.

In this paper, we will introduce *CLARIAH chaining search*: a platform to combine heterogeneous linguistic resources, specifically resources from the CLARIN ecosystem, and perform quantitative analyses in a single interface, while requiring only limited technical knowledge. This approach fits into the wider agenda of CLARIAH¹, the digital research infrastructure for arts and humanities in The Netherlands. CLARIAH aims to make a large set of linguistic resources available for integrated content search. CLARIAH distinguishes three levels of searchability. *Local searchability*, where the resource is available as a web service. *Federated content search*, where resources of the same type can be queried as a single resource. *Chaining search* is the most complex level in this framework: information from heterogeneous sources can be combined, and sequential search workflows can be executed.

2 Method

2.1 Approach

We chose to implement our chaining search platform as a Jupyter notebook (Thomas et al., 2016) and a Python library². Our Python library `chaininglib` uses `pandas` (McKinney, 2011) as the data structure that represents results from a query.

In an earlier step in the development process, we contemplated implementing chaining search by allowing users to build one large query in a web interface, which jointly accesses all resources. This however requires transforming all resources in a format that could be queried with the same query format³. Furthermore, users who want advanced functionality, would have had to learn how to build a highly complex query. Therefore, we came up with the idea of performing separate, smaller queries to the different

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

¹<https://www.clariah.nl/en/>.

²The library and notebook are available from: <http://github.com/INL/chaining-search>.

³The CLARIAH data infrastructure is based on Linked Open Data (LOD), which is more natural for some resources (lexica) than for others (e.g. corpora).

resources, and chaining them using Python program code⁴. The Jupyter notebook makes the chaining process transparent and configurable for the user.

2.2 Linguistic resources

We will now give an overview of the linguistic resources available by default in CLARIAH chaining search. This gives a perspective on the possibilities of combining resources. However, our framework is not limited to these resources: if required, the user can relatively easily add more resources.

Corpora Our chaining search application supports corpora that are based on the BlackLab corpus search engine⁵ (De Does et al., 2017) and the CLARIN Federated Content Search specification (FCS) (Stehouwer et al., 2012). Corpora using one of these protocols can easily be added, by appending their URL to a configuration file. Support of the FCS protocol opens up the possibility to add a multitude (currently more than 1200) of freely available corpora, all over Europe⁶. Support of the BlackLab corpus search engine enables more advanced search options, like metadata search. In our application, a number of contemporary and historical Dutch corpora are available by default.

Lexica Lexica represented as Linked Open Data according to the *Ontolex* Lexicon model (McCrae et al., 2017) are accessible in our application. A collection of contemporary and historical Dutch lexica are available by default in the application. More lexica can be added by writing a custom SPARQL query. In addition, we support accessing lexica hosted at our institute via the INT *lexicon service* API⁷.

Treebanks Supported treebanks are the Corpus of Spoken Dutch (CGN) and Lassy, a corpus of written Dutch, which can be queried via XPath queries.

2.3 Library architecture

The library `chaininglib` consists of 4 packages: `search`, `process`, `ui` and `utils`. These four packages together provide the possibility to search linguistic resources, combine or process them, and interact with the Jupyter notebook interface.

As a data structure for storing search results, we use Pandas DataFrames, table-like structures, with built-in processing methods. This is very convenient: functions expect DataFrames and yield DataFrames, so no time is lost converting data to the right format. The built-in methods already provide rich possibilities for processing, such as computing means, filtering, grouping and combining. Only the processing methods specifically needed for our purposes have to be implemented in our library, and our methods can use the built-in methods as primitives.

Searching `search` provides functions with which corpora, lexica and/or treebanks can easily be searched for a given word or pattern. The different resources can be searched in an analogous way:

```
results = create_lexicon(lexicon_name).lemma(word).search()
results = create_corpus(corpus_name).pattern(pattern).search()
results = create_treebank(treebank_name).pattern(pattern).search()
```

Our search operation will provide the user with a results object containing a Pandas DataFrame of so-called *keywords in context* (*kwic*), which can be extracted with the following call:

```
df = results.kwic()
```

Processing The `process` and `utils` packages handle processing of results. The `utils` package provides functions for general operations to be applied to the search results, like filtering or computing differences. The `process` part deals with linguistic processing of the data: extraction of a lexicon from a corpus, computing word frequencies, etc. For example, finding differences between the lemmata of a

⁴Cf. Verborgh et al. (2016), which advocates combination of resources in the client in a LOD context.

⁵<http://inl.github.io/BlackLab/>.

⁶See <https://www.clarin.eu/blog/clarin-federated-content-search-version-2> for CLARIN federated content search.

⁷<https://ivdnt.org/landingspagina-historisch-nederlands-lexica>

given result set with the lemmata of another results set (stored in DataFrames *df1* and *df2* respectively), is achieved with the following code:

```
diff = column_difference( df1["lemma_0"], df2["lemma_0"] )
```

This will yield a list of lemmata occurring only in DataFrame *df1*, a list of lemmata occurring only in DataFrame *df2*, and a list of lemmata occurring in both. The lexicon extraction function, mentioned before, takes a corpus DataFrame as an argument and yields a lexicon, once again stored in a DataFrame:

```
df_lexicon = extract_lexicon(df_corpus)
```

UI The `ui` package caters for the interaction between the library and the Jupyter notebook interface. This package makes it possible to create a search interface, with GUI controls, in the notebook, and to save and load DataFrames from the notebook.

2.4 Applications

The `chaininglib` library described in the previous section enables the user to perform all kinds of search and processing operations and, most importantly, chain them. The example notebook⁸ contains several different use cases. We will give two examples: one using a single resource, and one in which multiple resources are chained.

Single resource In some southern or south-western Dutch dialects, one tends to omit the phoneme *h* word-initially. That means that one might for instance find *and* instead of *hand*. In the chaining search application, this phenomenon can be investigated⁹ by gathering sentences with occurrences of lemmata starting with *h*, realized with or without *h*-dropping:

```
pattern_with_h = r'[lemma="h[aeo].*"&_word="h[aeo].*"'
pattern_without_h = r'[lemma="h[aeo].*"&_word="[aeo].*"'
```

We search for both patterns, separately count the number of occurrences for each location and show the results:

```
df = create_corpus(sailing_letters).pattern(pattern).search().kwic()
results_per_location = df[['lemma_0', 'location']].groupby('location').count()
display_df( results_per_location.sort_values(by=['lemma_0']), mode='chart' )
```

Chaining The second example investigates another phenomenon: in Dutch, male and female attributive adjectives get an *-e* suffix when used with indefinite determiner *een*, but sometimes they do not. To discover which adjectives allow this phenomenon to take place, we will chain different resources, using one to filter the other. We will first gather occurrences of attributive adjectives lacking the expected *-e* inflection from the CHN corpus (modern Dutch):

```
df_corp = create_corpus("openchn").pattern(' [pos="DET"&lemma="een"
[word=".*[^e]$"&_pos="AA.*degree=pos.*'] [pos="NOU.*gender=[fm].*"]' ).search().kwic()
()
```

Now we need to compare these *e*-less words from the corpus to a modern Dutch lexicon¹⁰. We look for adjectives without a final *-e* in their lemma, but which can get a final *-e* in attributive use. This must be done in two steps. We first search the lexicon for adjectives of which the lemma does not end with *-e*. Then we check if adjectives of this list are able to get a final *-e* in the wordform (because there are some known exceptions, which never get a final *-e*):

```
df_lex = create_lexicon("molex").lemma(' (.+)[^e]$' ).pos(' ADJ(degree=pos)' ).search().kwic()
final_e_condition = df_filter(df_lex["wordform"], 'e$')
df_lexicon_form_e = df_lex[ final_e_condition ]
```

⁸<https://github.com/INL/chaining-search/blob/master/Examples.ipynb>.

⁹Using the *Letters as Loot* corpus (Van der Wal et al., 2012).

¹⁰GiGaNt-Molex, <https://ivdnt.org/downloads/taalmaterialen/tstc-gigant-molex>

Now, we can filter the corpus results (from the very first step) using our lexicon results. When the lexicon tells us an adjective should have had a final *-e*, but its attributive occurrence in the corpus is spelled like its lemma (so it is not inflected), we keep this occurrence. The rest is filtered out. In this way, we get a list of sentences in which the phenomenon under investigation occurs:

```
e_forms = set(df_lexicon_form_e.lemma)
cond = df_filter(df_corp["word_1"], pattern=e_forms, method="isin")
result_df = df_corp[ cond ]
```

3 Case study: Compare sailing letters to lexicon

We apply CLARIAH chaining search to investigate how language was used by different social groups and by male and female speakers in the 17th and 18th century. As linguistic resource, we use the *Letters as Loot* corpus (Van der Wal et al., 2012). We filter and subsequently compare results based on social class of the sender, gender and century. An interesting finding is that female writers tend to use more words that signify personal topics (e.g. *child*, *goodnight*), whereas men discuss more business topics (e.g. *cargo*, *sugar*). This effect is stronger in the higher social class than in the lower class.

4 Conclusion and discussion

We have shown that a Python library, combined with a Jupyter notebook web interface, can open up new possibilities for linguistic research: it becomes easier to combine and process heterogeneous linguistic resources. Our case study on a corpus of sailing letters, combined with a lexicon, shows that this ease of processing can be helpful in linguistic research.

A Jupyter notebook seems to be a good solution to combine results from different linguistic resources, for users with limited programming skills. For users without any programming skills, a more user-friendly user interface may be needed, although this offers less flexibility. For heavy queries with long computation times, it may be better not to use the Jupyter notebook, but invoke `chaininglib` directly from a script. At the moment, `chaininglib` is not yet able to handle result sets larger than computer memory. Optimizations to handle large results sets may be added in the future.

References

- Jesse De Does, Jan Niestadt, and Katrien Depuydt. 2017. Creating Research Environments with BlackLab. In Utrecht University, NL and Jan Odijk, editors, *CLARIN in the Low Countries*, pages 245–257. Ubiquity Press, December.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolx-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Wes McKinney. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for High Performance and Scientific Computing*, 14:9.
- Herman Stehouwer, Matej Durco, Eric Auer, and Daan Broeder. 2012. Federated Search: Towards a Common Search Infrastructure. *LREC 2012*, page 5.
- Kluyver Thomas, Ragan-Kelley Benjamin, Prez Fernando, Granger Brian, Bussonnier Matthias, Frederic Jonathan, Kelley Kyle, Hamrick Jessica, Grout Jason, Corlay Sylvain, Ivanov Paul, Avila Damin, Abdalla Safia, Willing Carol, and Jupyter Development Team. 2016. Jupyter Notebooks - a publishing format for reproducible computational workflows. *Stand Alone*, pages 87–90.
- Marijke J. Van der Wal, Gijsbert Rutten, and Tanja Simons. 2012. Letters as loot: Confiscated Letters filling major gaps in the History of Dutch. In Marina Dossena and Gabriella Del Lungo Camiciotti, editors, *Pragmatics & Beyond New Series*, volume 218, pages 139–162. John Benjamins Publishing Company, Amsterdam.
- Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert. 2016. Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. *Journal of Web Semantics*, 37-38:184–206, March.

Manually PoS tagged corpora in the CLARIN infrastructure

Tomaz Erjavec¹

Jakob Lenardič²

Darja Fišer^{2,1}

¹Dept. of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

²Dept. of Translation, Faculty of Arts
University of Ljubljana
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

1 Introduction

This paper provides a comparison of corpora that are manually annotated for word-level morphosyntactic information, i.e. part-of-speech (PoS) tags, and are available for download within the CLARIN infrastructure. Since such corpora provide gold-standard data, they are an important resource for training new PoS taggers as well testing the accuracy of the existing ones. It is therefore valuable to have a better understanding of the languages that are supported in this way through CLARIN, under what licences such corpora are available for download and to compare their encodings and PoS tagsets used in order to see to what extent they are interoperable.

The rest of the paper is structured as follows: Section 2 gives an overview of the manually PoS tagged corpora available through the CLARIN infrastructure and compares their encodings and PoS tagsets; Section 3 compares the corpora against the most comprehensive multilingual dataset of PoS annotated corpora, namely the Universal Dependencies treebanks; and Section 4 concludes the paper.

2 The corpora

Within the CLARIN Resource Families initiative¹ (morphosyntactic tagging, lemmatisation, syntactic parsing, named entities, sentiment labels, etc.).² In this paper we focus on a subset of 14 corpora which have been manually tagged for word-level morphosyntactic information, commonly known as part-of-speech tags. We here exclude the overviewed treebanks, which, although including PoS tags as well, are typically smaller than PoS-only annotated corpora. Table 1 gives an overview of the corpora in question, listing some of their most relevant metadata, i.e. language, size, tagset, and licence.³

The first important aspect is the uneven distribution of languages, where Western European languages are much less represented than the Eastern and Central European ones. This can be attributed to two factors. First, due to legacy or other factors, not all existing Western European PoS tagged corpora are included in the VLO or national repositories, which was a condition for including them in the Key Resource Family. Second, and somewhat more tentatively, manually PoS annotated corpora are expensive to produce, and it is likely that authors of (large) Western European language corpora consider them too valuable to make them freely available, whereas authors who produced the corpora of (smaller) Central and Eastern European languages are more willing to do so, as they are less likely to be able to sell them and because in this way they hope to strengthen the language technology state-of-the-art for their language – this is certainly the case for the Slovenian, Croatian and Serbian corpora listed in the table and available from the CLARIN.SI repository.

Second, the corpora are available under a multitude of licences, for the most part without having to sign a licence agreement, but mostly limited to non-commercial usage and with share-alike, which limits the uptake of the corpora for commercial purposes.

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

¹<https://www.clarin.eu/resource-families>, Fišer and Lenardič (2019) have prepared an overview of 74 corpora available through the CLARIN infrastructure that display varying levels of manual annotation (morp

²<https://office.clarin.eu/v/CE-2019-1384-Manually-annotated-corpora-report.pdf>

³Since the publication of the CLARIN Resource Families report, new versions of some of the corpora have been published, so the list of corpora has been updated for this paper.

Corpus	Language	k-tokens	Tagset	Licence
ssj500k 2.2	Slovenian	586	MULTEXT, UD	CC BY-NC-SA
Janes-Tag 2.0	non-standard	75	MULTEXT	CC BY-SA
hr500k 1.0	Croatian	500	MULTEXT, UD	CC BY-SA
ReLDI-NormTagNer-hr 2.0	non-standard	89	MULTEXT	CC BY
SETimes.SR 1.0	Serbian	87	MULTEXT, UD	CC BY-SA
ReLDI-NormTagNER-sr 2.0	non-standard	92	MULTEXT	CC BY
MDET	Estonian	513	MULTEXT	CLARIN ACA
Szeged Corpus 2.0	Hungarian	1,500	MULTEXT-like	NORED-NC-ND
MATAS	Lithuanian	1,600	Lith. PoS tagset	CLARIN ACA
NKJP1M	Polish	1,000	IPI PAN Tagset	GNU GPL
CINTIL	Portuguese	1,000	CINTIL PoS tagset	ELRA \$
BNC Sampler	English	2,000	CLAWS 7	BNC (NORED)
MULTEXT-East "1984"	Multi	1,000	MULTEXT	CC BY-NC-SA
xLiME Twitter Corpus	Multi	364	UD	MIT

Table 1: Manually PoS tagged corpora downloadable within the CLARIN infrastructure.

The most common tagset in Table 1 is MULTEXT, or, more precisely, MULTEXT-East (Erjavec, 2012), which is used for all the 8 South Slavic corpora, as well as the corpora of Estonian and Hungarian texts.⁴ The MULTEXT-East is a positional tagset, i.e. the first character in the tag gives the part-of-speech, with the other giving the value of the position-determined attribute. So, for example, the tag *Ncndl* means *Category = Noun, Type = common, Gender = neuter, Number = dual, Case = locativge*.

We exemplify the use of MSDs and the basic encoding of the first 6 corpora in Table 1 with a sentence taken from the Slovenian social media corpus *CMC training corpus Janes-Tag 2.0* (Erjavec et al., 2017):

```
(1) <s>
    <w lemma="ta" ana="#Pd-nsn">To</w><c> </c>
    <w lemma="danes" ana="#Rgp">danes</w><c> </c>
    <w lemma="biti" ana="#Va-r3p-n">so</w><c> </c>
    <w lemma="zgolj" ana="#Q">zgolj</w><c> </c>
    <w lemma="slab" ana="#Agpfpn">slabe</w><c> </c>
    <w lemma="igralka" ana="#Ncfpn">igralka</w>
    <pc ana="#Z">.</pc>
</s>
```

By contrast, the Polish *NKJP1M* corpus, a manually annotated 1-million word subset of the National Corpus of Polish (Przepiórkowski, 2010), defines PoS tags, although also using TEI, slightly differently:

```
(2) <!-- ofierze [72,7] -->
    <f name="interps">
    <fs type="lex" xml:id="morph_1.16.1-lex">
    <f name="base"><string>oflara</string></f>
    <f name="ctag"><symbol value="subst"/></f>
    <f name="msd">
    <vAlt>
    <symbol value="sg:dat:f" xml:id="morph_1.16.1.1-msd"/>
    <symbol value="sg:loc:f" xml:id="morph_1.16.1.2-msd"/>
    </vAlt>
    </f>
```

Here, the annotation of the noun *ofierze* “victim” is marked up as belonging to the category SUBST, taken from the *ctag* package⁵ while the other features are those defined for the Morfeusz SGJP morphological

⁴However, the Hungarian tagset is a slightly modified one from the official MULTEXT-East specifications.

⁵<https://github.com/universal-ctags/ctags>

analyzer⁶, which was used for the annotation of the corpus. It should be noted that the inflectional features are not fully disambiguated, as the encoding includes the alternation (syncretism) of the dative and locative case forms.

The Lithuanian corpus *MATAS* also employs fine-grained morphosyntactic annotation, but differs quite a bit in the way the attributes are defined, as it uses its own dedicated tagset tailored to Lithuanian (Daudaravičius et al., 2007):

```
(3) <word="griežlių" lemma="griežlė" type="dktv mot.gim dgsk K">
<space>
<word="gyvenamose" lemma="gyventi(-a,-o)"
type="dlv teig nesngr neveik.r esam.l neįvardž mot.gim dgsk Vt">
<space>
<word="vietose" lemma="vieta" type="dktv mot.gim dgsk Vt">
<sep=".">
```

It can be noted that here the names of the features are in Lithuanian, rather than English or Latin based, making it more difficult to interpret them by non-Lithuanian speakers.

Finally, the examples above also give some idea of how the corpora are encoded: as can be seen, all of them use XML, but each one making use of different elements and attributes, even in cases (as goes for Slovenian vs. Polish) where both use TEI P5. Nevertheless, the format conversion from any of the XML schemas presented to a common one should not be too difficult – something that does not hold for the PoS tagsets, however.

3 Universal Dependencies

For many years, *TreeTagger*⁷ (Schmid, 1995) was the dominant tool for multilingual PoS tagging, as it comes accompanied by pre-trained models for over 20 languages. Currently, the Universal Dependencies project offers the largest (100 treebanks in over 70 languages) multilingual manually annotated corpus (Nivre and others, 2018), i.e. treebank, where all the word-level morphosyntactic features are also manually validated. On the basis of this corpus, the tool *UD-Pipe*⁸ (Straka and Straková, 2017) was trained, which annotates texts in the UD languages for morphosyntactic features, lemmas, and dependency links and labels. As it covers many languages using a common annotation framework, it is fast becoming the annotation tool of choice for PoS tagging as well. This brings with it the question of whether the corpora as overviewed in the previous section are still relevant for PoS tagger training at all.

Language	PoS Corpus	kilo tokens	UD Treebank	Size
Slovenian	ssj500k 2.2	586	UD SSJ	14
Croatian	hr500k 1.0	500	UD SET	197
Estonian	MDET	513	UD EDT:	434
Hungarian	Szeged Corpus 2.0	1,500	UD Szeged	42
Lithuanian	MATAS	1,600	UD ALKSNIS + HSE	46
Polish	NKJP1M	1,000	UD LFG + SZ	214

Table 2: Comparison of manually PoS tagged corpora with their UD counterparts.

Table 2 gives a comparison of the sizes of the most relevant (largest, and for standard language) corpora from Table 1 with the UD treebanks for the same languages. As can be seen, the sizes of the treebanks are typically much smaller (the only exception being Estonian), and it can thus be rather confidently predicted that PoS taggers trained on the PoS tagged corpora will achieve greater accuracy than those trained on the UD corpora.

However, and as noted, most of the PoS tagged corpora use different tagsets, whereas UD ones use a harmonised set of features, making them more universally useful for this reason. It would therefore be

⁶<http://sgjp.pl/morfeusz/>

⁷<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

⁸<http://ufal.mff.cuni.cz/udpipe>

interesting to investigate the optimal conversion of the native PoS tagsets to the UD ones. Currently, to the best of our knowledge, the UD corpora were produced by converting existing treebanks to the UD feature-set with the help of manually specified rules or mapping tables, which even required additional manual annotation for conversion as was the case with the conversion of the Slovenian MULTEXT-East tagset to the UD one. However, work has been done on automatic mapping of language-specific PoS tags to the Universal PoS tagset (Zhang et al., 2012), so a similar approach could also be used for these cases.

4 Conclusion

The paper has presented the manually PoS tagged corpora available via the CLARIN infrastructure and made the case for utilising them to train (better) PoS taggers than are currently available. This approach would be esp. beneficial, if the corpus-specific tagset would be mapped to the UD morphological features, thus improving the accuracy of UD-Pipe or training other taggers to perform multilingual tagging with a harmonised PoS tagset. Manually PoS tagged gold-standard corpora are part of the basic language resource kits and it should therefore be of strategic importance for CLARIN to actively encourage the integration of the existing gold-standard corpora in the CLARIN infrastructure in order to expand the portfolio of the supported languages. Furthermore, quality and interoperability of these basic datasets should be promoted and supported through shared tasks, data camps and hackathons.

References

- [Daudaravičius et al.2007] Vidas Daudaravičius, Erika Rimkutė, and Andrius Utkas. 2007. Morphological annotation of the Lithuanian corpus. In Bruno Poulouen Ralf Steinberger Jakub Piskorski, Hristo Tanev, editor, *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007*, pages 94–99, Prague, Czech Republic.
- [Erjavec et al.2017] Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Katja Zupan. 2017. *CMC training corpus Janes-Tag 2.0*. Slovenian language resource repository CLARIN.SI. <http://hdl.handle.net/11356/1123>.
- [Erjavec2012] Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 1:131–142.
- [Fišer and Lenardič2019] Darja Fišer and Jakob Lenardič. 2019. Overview of manually annotated text corpora. CLARIN Office Document.
- [Nivre and others2018] Joakim Nivre et al. 2018. *Universal Dependencies 2.3*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <http://hdl.handle.net/11234/1-2895>.
- [Przepiórkowski2010] Adam Przepiórkowski. 2010. TEI P5 as an XML Standard for Treebank Encoding. *IEEE Transactions on Learning Technologies - TLT*. <http://nlp.ipipan.waw.pl/~adamp/Papers/2009-tlt-tei/tei.pdf>.
- [Schmid1995] Helmut Schmid. 1995. Improvements In Part-of-Speech Tagging With an Application to German. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.
- [Straka and Straková2017] Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.
- [Zhang et al.2012] Yuan Zhang, Roi Reichart, Regina Barzilay, and Amir Globerson. 2012. Learning to Map into a Universal POS Tagset. In *EMNLP-CoNLL*.

A Use Case for Open Linguistic Research Data in the CLARIN Infrastructure. The Open Access Database for Adjective-Adverb Interfaces in Romance.

Gerlinde Schneider

Centre for Information Modelling
University of Graz, Austria
gerlinde.schneider@uni-graz.at

Christopher Pollin

Centre for Information Modelling
University of Graz, Austria
christopher.pollin@uni-graz.at

Katharina Gerhalter

Department of Romance Studies
University of Graz, Austria
katharina.gerhalter@uni-graz.at

Martin Hummel

Department of Romance Studies
University of Graz, Austria
martin.hummel@uni-graz.at

Abstract

The AAIF project is establishing appropriate ways to make linguistic research data on Adjective-adverbs in Romance languages openly accessible and reusable. Special focus is set on adhering to the FAIR data principles. Using a project-specific annotation model, it annotates corpora of linguistic phenomena related to adjectives with adverbial functions in Romance languages. This paper documents the approaches we use to accomplish these goals. An important part of this is the use and provision of data via the formats and interfaces defined by the CLARIN infrastructure.

1 Introduction

In 2017, the project *Open Access Database Adjective-Adverb Interfaces in Romance* (AAIF) won its funding as part of the Open Research Data (ORD) pilot programme by the Austrian Science Fund (FWF). The aim of the programme is to create role models and to gain experience with open access to research data in order to promote Open Science for FWF projects. Requirements for funding were: Research data must (1) be published on the basis of the latest technical standards, (2) must be openly accessible (Open Access), (3) reproducible, (4) machine-readable, (5) citable, (6) have an open licence for unrestricted further use and (7) must be published in a registered repository.¹ The call explicitly mentioned the FAIR data principles, 15 guidelines which aim to promote the findability and (re-) usability as well as sustainability of research data. FORCE11, the committee behind the FAIR principles identifies, “discovery of, access to, integration and analysis of, task-appropriate scientific data” as the main challenges for data-intensive science² and name four qualities research data should possess: F: To be Findable, A: To be Accessible, I: To be Interoperable, R: To be Re-usable.(Wilkinson et al., 2016) AAIF was funded to (1) Provide research data from numerous previous and ongoing FWF-funded projects in the most open way possible (2) Provide a reusable infrastructure for the long term preservation and provision of linguistic research data.³ This paper intends to document the approaches we use to accomplish these goals. An important part of this is the use and provision of data via the CLARIN infrastructure.

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

¹<https://www.fwf.ac.at/en/news-and-media-relations/news/detail/nid/20151221-2158/>

²<https://www.force11.org/fairprinciples>

³<https://adjective-adverb.uni-graz.at/en/research/projects/open-access-database-2017-2019>

2 Project and Data

The data dealt with in the AAIF project was created in the course of several individual research projects on adjective-adverb interfaces in Romance.⁴ Adjective-adverb interfaces are manifestations of a linguistic phenomenon that refers to adjectives that take over an adverbial function. Examples from Spanish are adjective-adverbs such as *volar alto* 'to fly high' or *ver claro* 'to see clear', discourse markers such as *cierto* 'true' or 'right' and adverbial prepositional phrases like *de seguro* 'certainly'. To study the phenomenon, more than 27,000 examples in French, Italian, Portuguese and Spanish have so far been collected and manually annotated, with examples in Romanian being added to the corpus until the project's end in late 2019.⁵ Each example consists of its contextual text passage and a narrower phrase that is relevant for the linguistic analysis. The respective adverb and, depending on the research question, the corresponding verbs, prepositions, and the subject of the phrase were annotated at different levels of depth. Information on morphosyntactic structure, inflection, attribution target, semantic classification, reduplication and coordination as well as the lemma is annotated for the adverb (Gerhalter et al., 2018). Within the project, an annotation model as domain-specific RDFs ontology was developed on the basis of this annotation practice and serves as a basis for further tagging and as a common conceptual model within the research group. (Pollin et al., 2018) One goal of the project is to integrate these data to one comprehensive database. The annotation model as well as all annotated research data available in the database are as provided in different data formats and serializations under a free license and via standardized and reasonable interfaces. The following sections describe this in more detail.

3 Open Linguistic Research Data leveraging the CLARIN Infrastructure and beyond

3.1 Metadata

One of the key principles for the creation of FAIR research data is describing the data with rich, accurate and accessible metadata. With the Component Metadata Infrastructure (CMDI),⁶ CLARIN offers a framework to create and use self-defined metadata formats. (Goosen et al., 2015) After creating a CMDI profile the complete information about our resource encoding can be found in the CMDI Component registry⁷. We focus on the preparation of a comprehensive metadata profile, covering editorial, descriptive, administrative, analytical and statistical metadata as proposed in a best practice by Koeva et al. (2012). Providing the CMDI metadata allows for harvesting and making it discoverable via the CLARIN Virtual Language Observatory (VLO),⁸ (Van Uytvanck et al., 2012) and makes a major contribution to the findability of our resources.

3.2 Domain-relevant standards

Converting the data into the Text Corpus Format (TCF) makes it available for use in the Weblicht⁹ tool chain. Weblicht is a widely used framework to annotate, investigate and analyze linguistic corpora. Providing the data in the TCF Format allows for interoperability with the Weblicht Tools and is therefore important for the reuseability of data within the CLARIN community.

3.3 Linked Linguistic Data

Semantic Web respectively Linked Data formats are gaining more relevance in the field of linguistic resources where they promote findability and interoperability. Our data is offered as highly structured RDF stored in a Blazegraph triple store and is referenced to a domain-specific ontology. A registration of our data set in the Linguistic Linked Open Data Cloud¹⁰ connects our resource with other linguistic resources available as Linked Data. (McCrae et al., 2016) We also convert the data into a reduced form of the NLP Interchange Format (NIF) (Chiarcos/Fäth, 2017) to exchange annotations.

⁴<https://adjective-adverb.uni-graz.at/en/>

⁵<https://gams.uni-graz.at/aaif>

⁶<https://cmdi.clarin.eu>

⁷<https://catalog.clarin.eu/ds/ComponentRegistry>

⁸<https://vlo.clarin.eu>

⁹<http://weblicht.sfs.uni-tuebingen.de>

¹⁰<http://linguistic-lod.org>

4 Repository Infrastructure

All project data is stored in the GAMS¹¹ institutional data repository. GAMS is a Fedora Commons based, OAIS compliant digital asset management system which serves the purpose of management, publication and long-term archiving of humanities research data. Data that has been created in projects in the fields of digital scholarly editions, cultural heritage, and also linguistic research can be conveniently published and enriched with corresponding metadata. Particular focus is set on the persistent storage and reusability of resources. (Stigler/Steiner, 2018) The repository builds upon a web service-based (SOAP, REST), platform-independent and distributed system architecture and follows an XML-based content strategy. Since 2014, GAMS has been a certified trusted digital repository in accordance with the guidelines of the Data Seal of Approval, since 2019 it is certified with the CoreTrustSeal. It is registered with the Registry of research data repositories re3data.org. The GAMS repository forms the core of the efforts to establish a Clarin-B Centre at the University of Graz by 2019.

5 Preparing the Infrastructure for Linguistic Research Data

Within the repository, data assets are organised and stored using specific content models. These models are designed to meet the requirements for data maintenance and the attribution of metadata in different research domains. For instance, data originating from textual scholarship is managed using a specific TEI content model. This model not only contains the individual text resource as a datastream but also offers services for dissemination in relevant output formats such as HTML or RDF and via defined APIs.

For the appropriate longterm-preservation and provision of linguistic research data—such as the data created within the AAIF project—the TEI content model is extended and also a specific container model for linguistic data and language resources is created. These offer corpus relevant methods, formats and programming interfaces such as those defined and offered by the CLARIN ERIC in a generic way.

In particular, this means providing the metadata via a CMDI interface and thus ensuring discoverability and reuse as well as providing the data in the TCF Format for interoperability with the Weblicht Tools. From a technical point of view the existing infrastructure is adapted and extended to match CLARIN's requirements, which covers the adaption the existing OAI-PMH endpoint for harvesting of CMDI metadata for findability in the VLO and the implementation of disseminators (access methods) for component based metadata for the linguistic object models. Further, it covers the implementation of conversion strategies from the present data format (mostly TEI/XML) into the TCF format as well as other relevant formats for the processing of language data such as the vertical format or plain text. This generic solution allows us to prepare metadata from other language resources available in the repository for its publication in the VLO and its data to be (re-)used with little effort in new research contexts.

6 Conclusion

De Jong et al. (2018) point out that CLARIN's strategy and vision have already been consistent with the later introduced FAIR principles from the very beginning. Our experience has shown that the VLO and its indexing of comprehensive and aligned metadata play an essential role when it comes to resource discovery and findability as well as access. Providing data, using the services of the CLARIN infrastructure, respectively of one of the CLARIN Centres, therefore contributes significantly to the FAIRness of linguistic research data. Not to be neglected, however, are the possibilities Linked Data offers us, in terms of interoperability. This certainly represents an aspect for future research and development.

¹¹<https://gams.uni-graz.at>

References

- Chiarcos, C. and C. Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. *Language, Data, and Knowledge. LDK 2017*, Lecture Notes in Computer Science, 10318, pp. 74-88. https://doi.org/10.1007/978-3-319-59888-8_6.
- De Jong, F., et al. 2018. CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 3259-3264. <https://dspace.library.uu.nl/handle/1874/364776>
- Gerhalter K., et al. 2018. Compilation and Annotation of Adjective-Adverb Interfaces in Romance. Towards a multilingual Open Access Corpus. *CHIMERA: Romance Corpora and Linguistic Studies*, 5(2), pp. 115-121. <http://dx.doi.org/10.15366/chimera2018.5.2.009>
- Goosen T., et al. 2015. CMDI 1. 2: Improvements in the CLARIN Component Metadata Infrastructure. *Selected papers from the CLARIN 2014 Conference*, pp. 36-53. <https://hdl.handle.net/20.500.11755/91536b93-31cb-4f4a-8125-56f4fe0a1881>
- Koeva, S., et al. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, Vol. 0, No. 1, pp. 65-110. <http://dx.doi.org/10.15398/jlm.v0i1.33>
- McCrae, J. P., et al. 2016. The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2435-2441. http://www.lrec-conf.org/proceedings/lrec2016/pdf/851_Paper.pdf.
- Pollin, C., et al. 2018. Semantic Annotation in the Project "Open Access Database Adjective-Adverb Interfaces in Romance". *Proceedings of the Workshop on Annotation in Digital Humanities. CEUR Workshop Proceedings*, pp. 41-46. ceur-ws.org/Vol-2155/pollin.pdf
- Stigler, J. H., Steiner, E. 2018. GAMS - An Infrastructure for the Long-term Preservation and Publication of Research Data from the Humanities. *Mitteilungen der Vereinigung Oesterreichischer Bibliothekarinnen und Bibliothekare*, 71(1), pp. 207-216. <https://doi.org/10.31263/voebm.v71i1.1992>
- Van Uytvanck, D., et al. 2012. Semantic metadata mapping in practice: the Virtual Language Observatory. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 1029-1034. http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf
- Wilkinson, M. D., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3. <https://doi.org/10.1038/sdata.2016.18>

CLARIN Web Services for TEI-annotated Transcripts of Spoken Language

Bernhard Fisseni, Thomas Schmidt

Leibniz-Institut für Deutsche Sprache (IDS)
Mannheim, Germany

{fisseni, thomas.schmidt}@ids-mannheim.de

Abstract

We present web services implementing a workflow for transcripts of spoken language following TEI guidelines, in particular ISO 24624:2016 “Language resource management – Transcription of spoken language”. The web services are available at our website and will be available via the CLARIN infrastructure, including the Virtual Language Observatory and WebLicht.

1 Introduction / Recapitulation

Schmidt et al. (2017) sketch, and partly implement, an architecture for making CLARIN webservices usable for transcriptions of spoken language, focusing on the TEI-based standard ISO 24624:2016 “Language resource management – Transcription of spoken language” [henceforth **ISO/TEI**] as a pivot format for web services. The 2017 paper concentrates on a solution with an encoder/decoder pair which first transforms ISO/TEI to WebLicht’s TCF¹ and re-transforms the TCF result of the service chain to ISO/TEI. Thus a large class of language technology tools becomes accessible to researchers working with spoken language while maintaining interoperability with tools which are commonly used for manual transcription and annotation of audiovisual material (such as ELAN, Praat or EXMARaLDA).

As Schmidt et al. (2017) argue, CLARIN’s service-oriented approach could be further leveraged for spoken language data through the development of services which (a) take into account the specific characteristics of spoken data as well as the specific tasks arising in their curation, and which (b) operate directly on the ISO format without a ‘detour’ via TCF. The present contribution explores this option further with respect to a typical use case: curation of interview data (sec. 3), sketching a workflow for related use cases and describing a CLARIN-conformant implementation of this workflow (sec. 4).

2 Related Work

Workflows for the curation of interview data have been discussed in the CLARIN context on the occasion of several workshops on Oral History (<https://oralhistory.eu/>) where the focus of this work was on the use of speech technology (e.g. ASR, forced alignment) rather than on enriching textual transcription data with language technology, as in the current paper. Ideally, both approaches complement each other.

Several methods described here were originally developed in the context of the EXMARaLDA system (<http://www.exmaralda.org>), as part of the workflow for compiling the Research and Teaching Corpus of Spoken German (FOLK, see Schmidt 2016) and/or as components of curation workflows at the CLARIN B-centres Hamburg Center for Language Corpora (HZSK, <https://corpora.uni-hamburg.de/>) and the Archive for Spoken German at IDS (AGD, <http://agd.ids-mannheim.de/>). Details on the development of the POS tagging model are described by Westpfahl (2019). Several of the services described in sections 4 reuse and extend these methods (at least conceptually) and put them on a different technological basis thereby integrating them more fully into the CLARIN infrastructure.

Besides Schmidt et al. (2017) and the ISO specification itself, the role of TEI as a suitable basis of a standard for spoken language transcription has been discussed, among others, by Schmidt (2011) and Liégeois et al. (2017). The TEI guidelines’ chapter 8 on “Transcriptions of Speech” has also been used in CLARIN resources such as the GOS Corpus of Spoken Slovene (see Verdonik et al. 2013) and as the basis for a CLARIN-wide format for parliamentary data.

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

¹see https://weblight.sfs.uni-tuebingen.de/weblightwiki/index.php/The_TCF_Format

3 Use case: Legacy interview corpora

AGD hosts more than 80 spoken language corpora with more than 10,000 hours of audio or video recordings, increasing continuously through external collaborations and data deposits. AGD curates such external resources, i.e. putting audio/video recordings, metadata, transcripts and annotations into a state where they can be archived, discovered (= found) and reused. The main datatypes in AGD are *interaction corpora* (e.g. the FOLK corpus, Schmidt 2016), *variation corpora* (e.g. *Deutsch Heute* or *Australiendeutsch*), and *interview corpora*, on which we would like to focus in the present paper. An example of a curated interview corpus at AGD is Norbert Dittmar's *Berliner Wendekorpus*². Currently under curation are, e.g., the audio recordings from an interview study on German refugees in Britain ("Kindertransporte", see Thüne 2019). Multilingualism plays a central role for these data because speakers have migration histories and recordings hence often include code switching or mixing. These data share a high potential for interdisciplinary reuse, e.g. in sociological or oral history studies, and similar data is curated in many other centres (mostly outside CLARIN).

Typical initial data deposits consist of audio, transcripts in modified orthography (in English, e.g., "dunno" for "don't know") in some word processor format, and more or less structured metadata in legacy formats. AGD curates such data (a) fully digitising the resource, (b) transforming textual data into structured, interoperable formats conforming to best practices and standards, (c) interconnecting different data types (e.g. aligning transcripts with recordings), (d) enriching data with linguistic information (e.g. POS-tagging), and (e) integrating them into DGD and into the wider language resource infrastructure (e.g. assigning PIDs, offering OAI/PMH). Curation workflows have common building blocks, which we propose to implement in a set of ISO/TEI-based, CLARIN-conformant web services. The same methods and tools can be used in a much wider range of contexts than just the specific use case illustrated here.

We use the following example³ from the *Corpus Australian German* throughout the text, and show excerpts from the results of steps in the toolchain.

```
MC: Welche Früchte ham sie (.) hier in der (-) Gegend?
AS: Äh, Apfel. [...]
MC: Und ähm vielleicht könnten wir n bisschen umschalten ins Englische.
    What part of Germany did your forefathers come from?
AS: Eh, our people came from what they call Schlesien.
    I wouldn't know how you pronounce that in English.
```

4 Workflow and Tools

We provide an abstract description of the functionality of the services and an explanation of the motivation and challenges for each step.⁴ The process is conceived of as a pipeline, so that the output of one step can immediately serve as input to the next step. We will also mention some parameters, but have to refer the reader to the documentation for a detailed description. All services can be given a default language which will be used if there isn't a more specific language annotation in the documents, or the language cannot be detected. Contrary to the approach in TCF, ISO/TEI documents thus inherently support multilingual texts.

4.1 Plain text to ISO/TEI-annotated texts (text2iso)

As detailed above, our use case regarding legacy corpora starts with documents in word processor format. We can disregard most of the formatting and expect input in plain text format for our web services. Hence the first step is to convert plain text transcribed data to a ISO/TEI-conformant format, which serves as input for all further processing steps.

In this step, the main challenge is to specify a plain text input format that is sufficiently expressive to serve in the most common cases, and sufficiently simple and restricted to be typed and parsed efficiently; parsing errors and other difficulties are signalled in XML comments. The format is supposed to allow segmentation of the conversation into utterances, and assignment of these utterances to speakers. A specification is available at <https://github.com/Exmaralda-0rg/teispeechtools/blob/master/doc/Simple-EXMARaLDA.md>. The result of this step is a transcription file which is split into utterances: an `<annotationBlock>` for each utterance contains a `<u>` element as well as `<incident>` elements containing non-verbal actions and `<spanGrp>` elements containing commentaries. A `<timeline>` is derived from the text, and all annotation is situated

²<http://hdl.handle.net/10932/00-0332-BD7C-3EF5-0B01-4>, http://agd.ids-mannheim.de/BW--_extern.shtml

³http://hdl.handle.net/10932/00-0332-BCFF-D7B3-7A01-9,AD--_E_00010

⁴The web services are available at <http://clarin.ids-mannheim.de/teilicht>. The functionality is also available as a Java library and command-line tool, see <https://github.com/Exmaralda-0rg/teispeechtools/>.

with respect to the <timeline>. Elements of the timeline are the beginning and end of each utterance; in case of overlap, the overlap start and end is referenced as an <anchor> within the utterances.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader>
  <profileDesc><particDesc>
    <person n="AS" xml:id="AS"><persName><abbr>AS</abbr></persName></person>
    <person n="MC" xml:id="MC"><persName><abbr>MC</abbr></persName></person>
  </particDesc></profileDesc> <encodingDesc>...</encodingDesc> <revisionDesc>...</revisionDesc>
</teiHeader> <text xml:lang="de">
  <timeline unit="ORDER">
    <when xml:id="B_1"/> <when xml:id="E_1"/> <when xml:id="B_2"/> <when xml:id="E_2"/>
    <when xml:id="B_3"/> <when xml:id="E_3"/> <when xml:id="B_4"/> <when xml:id="E_4"/>
    <when xml:id="B_5"/> <when xml:id="E_5"/> <when xml:id="B_6"/> <when xml:id="E_6"/>
    <when xml:id="B_7"/> <when xml:id="E_7"/> <when xml:id="B_8"/> <when xml:id="E_8"/>
  </timeline>
  <body>
    <annotationBlock start="B_1" end="E_1" who="MC">
      <u>Welche Früchte ham sie (.) hier in der (..) Gegend?</u>
    </annotationBlock> <annotationBlock start="B_2" end="E_2" start="B_2" who="AS">
      <u>Äh, Apfel.</u>
    </annotationBlock> ... <annotationBlock start="B_8" end="E_8" who="AS">
      <u>I wouldn't know how you pronounce that in English.</u>
    </annotationBlock>
  </body>
</text>
</TEI>
```

4.2 Segmentation according to transcription convention (segmentize)

In the next step, the text is segmented according to transcription conventions. We enforce a tokenisation into words in <w> elements and punctuation in <pc>, and some information is lifted from the plain text of an <u> to the annotation level, mainly pauses and unclear or incomprehensible text. ISO/TEI allows to use time <anchor> elements also in the middle of words. Keeping the <anchor>s in place while processing the surrounding plain text was one of the challenges of implementing this step, as in this case, XML structure interferes with the abstract structure of the transcription.

```
<annotationBlock start="B_1" end="E_1" who="MC"><u>
  <w>Welche</w> <w>Früchte</w> <w>ham</w> <w>sie</w> <pause type="micro"/>
  <w>hier</w> <w>in</w> <w>der</w> <pause type="short"/> <w>Gegend</w> <pc>?</pc>
</u></annotationBlock>
```

4.3 Language detection (guess)

The motivation for this step is that interview data are often massively multilingual, and it is useful to be able to assign languages to single utterances. In contrast to TCF, the TEI formats allow @xml:lang on every structural level of text.

The service uses the Apache OpenNLP (<https://opennlp.apache.org/>) language models and language detector to process single utterances and guess what language they are in. It is possible to constrain the search space to a set of languages to avoid mis-detection of similar languages like German and Low German. A configurable threshold (default: 5 words) can be set to prevent language detection in utterances that are too short for a reliable result. In the result, the <u> have been annotated with @xml:lang attributes.

```
<annotationBlock start="B_5" end="E_5" who="MC"><u xml:lang="de">
  <w>Und</w> <w>ähm</w> <w>vielleicht</w> <w>könnten</w> <w>wir</w> ... </u>
</annotationBlock> <annotationBlock start="B_6" end="E_6" who="MC"><u xml:lang="en">
  <w>What</w> <w>part</w> <w>of</w> <w>Germany</w> <w>did</w> ... </u>
</annotationBlock>
```

4.4 OrthoNormal-like Normalisation (normalize)

EXMARaLDA includes the OrthoNormal tool for transcript normalisation, i.e. the mapping of tokens in modified orthography to their standard orthographic equivalent. The automated part of normalisation is dictionary-based and only available for German at the moment (see Schmidt 2012). Normalisation works on the <w> elements, which are annotated with a @norm attribute containing the normalised form.

```
<annotationBlock start="B_1" end="E_1" who="MC"><u xml:lang="de">
  <w norm="welche">Welche</w> <w norm="Früchte">Früchte</w> <w norm="haben">ham</w>
  <w norm="sie">sie</w> <pause type="micro"/> <w norm="hier">hier</w> <w norm="in">in</w>
  <w norm="der">der</w> <pause type="short"/> <w norm="Gegend">Gegend</w> <pc>?</pc>
</u></annotationBlock>
```

4.5 POS-Tagging with the TreeTagger (pos)

Preferrably after normalisation, a document can be part-of-speech-tagged and lemmatised. Tagging is done using TreeTagger via TT4J.⁵ We use the standard tagging models provided by the TreeTagger, which are mainly for written language but include a model for spoken French, and additionally a model trained on spoken German by Westpfahl (2019). Respecting the language of the current word <w>, the correct parser model is chosen by language, and the @pos and @lemma attributes are set accordingly.

```
<annotationBlock start="B_5" end="E_5" who="MC"><u xml:lang="de">
  <w lemma="und" norm="und" pos="KON">Und</w> ... <w lemma="in" norm="ins" pos="APPRART">ins</w>
  <w lemma="Englische" norm="englische" pos="NN">Englische</w> <pc>.</pc>
</u></annotationBlock> <annotationBlock start="B_6" end="E_6" who="MC"><u xml:lang="en">
  <w lemma="what" pos="DTQ">What</w> ... <w lemma="come" pos="VVB">come</w> <w lemma="from" pos="PRP">from</w>
  <pc>?</pc>
</u></annotationBlock>
```

4.6 Pseudo-alignment using Phonetic Transcription or Orthographic Information (align)

Another addition to the EXMARaLDA workflow is pseudo-alignment between transcription and recordings using graphemic or phonemic information. Most of the data submitted to the paradigmatic workflow are not aligned, i.e. do not contain timestamps pointing from the transcript to the recording. A logical approach is to apply *forced alignment* on these. Several aligners exist, most importantly MAUS, provided by the Bavarian Archive for Speech Signals (BAS).⁶ If possible, we use MAUS in our workflow. However, poor quality of the audio, large file sizes, or legal restriction can make this difficult or impossible. For such cases, a pseudo-alignment, which estimates an alignment based on the graph(em)ic form of utterances, i.e., counting letters or phone(me)s derived from a grapheme-to-phoneme conversion, is a useful alternative. Optionally, the canonical phonetic transcription can be added to the ISO/TEI document using the attribute @phon on <w> elements.

The alignment thus achieved can be manually improved, if necessary.

```
<timeline><tei:when interval="0s" xml:id="B_1"/> <tei:when xml:id="E_1" interval="5.394s" since="B_1"/>
<tei:when xml:id="B_2" interval="5.394s" since="B_1"/> <tei:when xml:id="E_2" interval="6.356" since="B_1"/> ...
</timeline> <body>
  <annotationBlock end="E_2" start="B_2" who="AS"><u start="B_2" end="E_2">
    <w lemma="Äh" norm="äh" phon="ʔɛ:" pos="ADJA">Äh</w> <pc>,</pc>
    <w lemma="Apfel" norm="Apfel" phon="ʔap.fəl" pos="NN">Apfel</w> <pc>.</pc>
  </u></annotationBlock> ...
```

5 Conclusion and Outlook

We hope to have shown that web services centred around ISO 24624:2016 form a useful addition to the CLARIN universe. The web services are currently available from IDS and will have been fully integrated into the CLARIN infrastructure by the time of the conference. Once the base architecture is thus established, we see several ways of evaluating and improving the individual services, e.g. by offering a direct choice of the tagger models for specific languages, or by testing whether language detection with moving windows can be applied to longer utterances in a way that detects language shifts like code switching.

References

- Liégeois, L., Benzitoun, C., Etienne, C., & Parisse, C. (2017). Vers un format pivot commun pour la mutualisation, l'échange et l'analyse des corpus oraux. In *FLORAL*. Orléans, France.
- Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the TEI*, 1, 1–22.
- Schmidt, T. (2012). EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In T. Declerck, K. Choukri, & N. Calzolari (Eds.), *Proceedings of LREC'12* (pp. 236–240). ELRA.
- Schmidt, T. (2016). Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for language technology and computational linguistics*, 31(1), 127–154.
- Schmidt, T., Hedeland, H., & Jettka, D. (2017). Conversion and annotation web services for spoken language data in CLARIN. In L. Borin (Ed.), *Selected papers from the CLARIN annual conf. 2016* (pp. 113–130). Linköping University Electronic Press.
- Thüne, E.-M. (2019). *Gerettet*. Berlin, Leipzig: Hentrich & Hentrich.
- Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47(4), 1031–1048.
- Westpfahl, S. (2019). *Dissertation (unpublished)* (PhD thesis). Universität Mannheim.

⁵see <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/> and <https://reckart.github.io/tt4j/>, respectively.

⁶<https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface>

Using DiaCollo for historical research

Bryan Jurish

Berlin-Brandenburgische Akademie der
Wissenschaften
Berlin, Germany
jurish@bbaw.de

Maret Nieländer

Georg-Eckert-Institut – Leibniz-Institut für
internationale Schulbuchforschung
Braunschweig, Germany
nielaender@leibniz-gei.de

Abstract

This article presents some applications of the open-source software tool DiaCollo for historical research. Developed in a cooperation between computational linguists and historians within the framework of CLARIN-D's discipline-specific working groups, DiaCollo can be used to explore and visualize diachronic collocation phenomena in large text corpora. In this paper, we briefly discuss the constitution and aims of the CLARIN-D discipline-specific working groups, and then introduce and demonstrate DiaCollo in more detail from a user perspective, providing concrete examples from the newspaper "*Die Grenzboten*" ("messengers from the borders") and other historical text corpora. Our goal is to demonstrate the utility of the software tool for historical research, and to raise awareness regarding the need for well-curated data and solutions for specific scientific interests.

1 Introduction

Ever since their establishment in 2011, German CLARIN centers have worked together with discipline-specific working groups to develop and improve their services in close dialogue with the needs of philologies, history, social science, etc. The German CLARIN initiative, CLARIN-D, has strong roots in computational linguistics. With the help of the working groups, it has been possible to curate and integrate corpus data that is important to different fields of the humanities and social sciences, as well as to disseminate knowledge of the usefulness of computational linguistic methods for other disciplines.

Developed in a collaboration between historians and computational linguistics within the context of the CLARIN-D working groups, DiaCollo (Jurish, 2015) is an open-source software tool for exploration and interactive visualization of diachronic change with respect to collocation behavior in large collections of (historical) text. In addition to the technical, implementation-oriented issues common to all software development projects on the one hand and the various challenges of source criticism characteristic for historical research on the other, interdisciplinary collaborations of this kind present challenges all their own, ranging from lack of established shared terminology (e.g. "term", "concept", "query", "type/token", "collocant/collocate", "relevance") to fundamentally different approaches to what constitutes "research activity" as such (analytic/stipulative vs. hermeneutic/interpretive). Over the course of the collaboration, DiaCollo underwent several iterations of the software development lifecycle phases of "planning", "implementation", and "evaluation" – in the latter case relying on extensive feedback from the working group's historians to identify missing functionality and potentially useful new features.

After its initial release, DiaCollo was integrated into the corpus administration framework of the CLARIN service center at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW). At the time of writing (April, 2019), DiaCollo indices for 70 distinct curated text corpora comprising a total of over 15,000,000,000 (15G) source tokens have been indexed and deployed at the BBAW, where they enjoy a modicum of popularity with an average of about 500 queries per day over the past

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

12 months. Of these curated corpora, 18 indices are publicly accessible and 3 more can be queried using CLARIN credentials.¹

2 Background

In linguistics, collocations are sets of words or terms that frequently occur in one another's vicinity, presumably because they belong to the same "semantic field" and thus shape their respective meanings, as suggested by J.R. Firth's well-known assertion that "you shall know a word by the company it keeps" and Wittgenstein's famous "*die Bedeutung eines Wortes ist sein Gebrauch in der Sprache*" ("the meaning of a word is its use in the language"). For example, the fact that the words "smoke" and "fire" tend to occur near one another in a text corpus suggests that there is indeed a semantic relation between them – in this case, a causal one.

Previous work in computational linguistics has established a number of methods for unsupervised discovery of collocations in text corpora, based on distributional properties of the collocated terms alone (see e.g. Evert, 2008). Informally, distributional collocation discovery procedures identify those word-pairs as potential collocations which occur together substantially more often than would be expected under "chance" conditions. Collocation profiling is a related technique which requires the user to provide one or more search terms of interest (the "collocant"), and searches for those terms in the corpus which associate most strongly with the collocant (i.e. the "collocates"). The association strength of a particular candidate collocate is estimated with regard to its own independent frequency in the corpus as well as that of the collocant, and should provide a quantitative approximation of the "relevance" of the respective collocate for the given collocant. To illustrate, a simple collocation profiling procedure would investigate all words in a pre-defined neighborhood of the search term, e.g. within a window of 5 words to the left and right. The more often one of these words occurs together with the collocant, compared to its frequency in the corpus overall, the stronger its association with the search term will be.

Synchronic collocation analysis has long been employed to provide evidence for typical usage(s) of words/concepts in the corpus as a whole, that is, semantics. It is also possible to compare collocation-profiles of different words, to look at differences and similarities in usage (e.g. for lexicography). "Ready-to-use" implementations include both the DWDS "*Wortprofil*" database² and Cyril Belica's co-occurrence database "CCDB"³. More complex user queries are possible (and familiarity with the associated software tools and interfaces required) when using the *Deutsches Referenzkorpus* (DeReKo) with the "COSMAS II" interface.

When analyzing historical text, synchronic collocation analysis can be a point of departure for comparing the usage of certain terms in historic source material with their use in the contemporary reference corpora. In order to be truly useful for historical research, collocation analysis should also provide methods that reveal changes in language use over time (in specific corpora), allowing users to trace phenomena such as semantic shifts, discourse trends, history of concepts, introduction of neologisms, etc. DiaCollo has been specifically developed for this purpose. As a free, open-source, language-agnostic software package⁴, it can also be integrated into other project contexts and corpus infrastructures.

DiaCollo corpus data must be pre-tokenized and each document must be assigned a characteristic date (e.g. year of publication) to represent the diachronic axis. If provided by the corpus, DiaCollo can also make use of additional token-level attributes such as lemmata or part-of-speech tags as well as document-level metadata such as author or genre to enable finer-grained queries and aggregation of result profiles (Jurish, 2018). As with any other data-driven procedure, DiaCollo is subject to "garbage-in / garbage-out" phenomena: "messy" corpora containing abundant OCR or annotation errors, mistokenizations, and/or incorrect document metadata are less likely to produce satisfying results for humanities researchers than "tidy", well-curated corpora with accurate metadata and reliable linguistic annotations.

¹ <https://kaskade.dwds.de/~jurish/diacollo/corpora/> (this and subsequent URLs last accessed 16 April, 2019)

² <https://www.dwds.de/d/ressourcen#wortprofil>

³ <http://corpora.ids-mannheim.de/ccdb/>

⁴ <https://metacpan.org/release/DiaColloDB/>

3 Use Case: Debates on Education in *Die Grenzboten*

As an introduction to DiaCollo's functionality, we will consider the collocates of a simple search term *Schule* ("school") in the largest historical corpus available at the BBAW, the *Deutsches Textarchiv*⁵ ("German Text Archive", DTA). Presentation of results in HTML format displays up to the specified number (kbest) of collocates (e.g. 10) for *Schule* discovered within the chosen time slice (e.g. a decade) in the form of a table. Each row of the table includes a color-code indicating the strength of the collocate's association preference as well as links to (close approximations of) the underlying corpus evidence for the corresponding collocation pair as Keywords-in-Context (KWIC), allowing the user to focus her attention more closely on the original text source. Additional visualization modes such as the "bubble" and "cloud" formats display changes in the collocates on an interactive timeline. For the example query, the collocates give quite obvious evidence that the term *Schule* is associated with words within the semantic field of the institution of the church in the earliest documents queried (e.g. in the 1560s: *Kloster* ("cloister"), *Pfarrherr* ("pastor"), and *Kirche* ("church")). The findings imply that the influence of this institution on the school system begins to mingle with worldly institutions in texts from the 1710s, where collocates include *Kirche* ("church"), as well as *Inspektor* ("inspector"), *preußisch* ("prussian"), and *Universität* ("university"); the term "church" disappears from the lists of top-10 collocates from the 1770s onwards (but re-occurs in the 1840s and 1890s).

We will now further demonstrate the use of DiaCollo by looking at German education policy as discussed in a historical periodical. *Die Grenzboten* was a German-language national-liberal magazine published from 1841 to 1922. Originally published as a (bi-)weekly periodical, the 311 volumes (roughly 180,000 pages) of *Die Grenzboten* were first digitized by the Staats- und Universitätsbibliothek Bremen⁶ with funding from the German Research Association (DFG), and have been integrated into the BBAW CLARIN service center's corpus infrastructure. *Die Grenzboten* covered a wide range of subjects in politics, literature, and the arts throughout the 'long' nineteenth century, and over the course of its publication was witness to several changes and attempted reforms of school systems in German-speaking territories. Using DiaCollo, we will now explore *Die Grenzboten's* stance on education policy.

3.1 Is the corpus a source for research into the history of education?

A time series analysis of the absolute frequency of selected relevant terms such as *Schule* ("school"), *Schulgesetz* ("school law"), *Schulbuch* ("textbook"), and other terms denoting various types of German schools shows that the lemma "school" was indeed mentioned in every year of *Die Grenzboten's* publication. Its raw frequency peaked at more than 500 tokens in 1890.⁷ Its relative frequency in the *Die Grenzboten*-corpus is twice as high⁸ as in the corresponding texts (1840–1920) from the aggregated DTA and *Digitales Wörterbuch der deutschen Sprache* (DWDS)⁹ "core" corpus. A DiaCollo search¹⁰ for collocates of *Schule* in ten-year epochs beginning at 1840 provides ample results from which to explore the school-related topics discussed in *Die Grenzboten*. Of the top-10 collocates per decade, most are nouns, some adjectives and one a finite verb (*gehören*, "to belong").

3.2 Are all findings relevant? Disambiguation by targeted close reading

DiaCollo's KWIC facility allows one to quickly check whether the results are applicable to a particular research question. In this case, strong adjective collocates of *Schule* are often associated with the sense of "school" as "doctrine", e.g. an artistic school or school of thought, which is irrelevant when looking at education policy. Another adjective collocate of interest is the lemma *hoch* ("high"). In DiaCollo's 'cloud' visualization for this query, it becomes evident that the adjective already appeared among the ten best collocates per epoch after 1870. Examination of the corresponding KWIC hits reveals that these collocates refer almost exclusively to secondary ("higher") schools, supporting the impression

⁵ <https://kaskade.dwds.de/~jurish/cac2019/Schule-dta>

⁶ University of Bremen: Grenzboten project, <https://www.suub.uni-bremen.de/ueber-uns/projekte/grenzboten/>

⁷ <https://kaskade.dwds.de/~jurish/cac2019/Schule-ts>

⁸ <https://kaskade.dwds.de/~jurish/cac2019/hist-gb>

⁹ <https://kaskade.dwds.de/~jurish/cac2019/hist-dta+dwds>

¹⁰ <https://kaskade.dwds.de/~jurish/cac2019/Schule-gb>

that *Die Grenzboten* was on the whole more concerned with higher education than with the *Volksschule* which provided basic primary (rural) education.

3.3 Do the findings offer tracks to specific discourses/debates?

Finally, the adjectives *konfessionell* (“denominational”) and *öffentlich* (“public”) were examined. These collocates appear among the top ten between 1860 and 1879, as do the nouns *Gemeinde* (“parish”/“congregation”) and *Kirche* (“church”) – the latter being as prominent and persistent as more expected noun collocates such as *Kind* (“child”) or *Lehrer* (“teacher”). Using DiaCollo’s on-the-fly filtering function to restrict our attention to adjective collocates only¹¹, the 1860s and 1870s documents reveal the adjectives *protestantisch* (“protestant”) and *evangelisch* (“evangelical”) as well as *katholisch* (“catholic”) as strong collocates of *Schule*.

We may assume that the prominence of this terminology involving religious denominations at that particular time was caused by the contemporary debates – since referred to as the *Kulturkampf* (“cultural struggle”) – concerning the rights and spheres of influence of state (Prussia) and church (Pope Pius IX) which started in some German territories in the 1860s and reached their peak in the 1870s. The debates involved the issue of who should be in charge of education and curricula, and how to deal with different religious denominations in schools. Loyal supporters of the Roman Catholic Church were referred to as *ultramontan* (“ultramontane”) during this period. A simple frequency query¹² shows that this kind of terminology is indeed present in the *Grenzboten* corpus, the former peaking and the latter beginning in the 1870s. Among the strong collocates of *Kulturkampf* and *ultramontan* are no terms that would hint at education though.

The connection only becomes clear if we turn our attention to all co-occurrences of *ultramontan* and GermaNet (Hamp & Feldweg, 1997) hyponyms of the synset *Bildungseinrichtung* (“educational institution”) or compounds matching a simple regular expression and using a rather broad paragraph-wide search window. Through closer reading of the corpus hits, we find evidence for anti-Catholic opinions in debates about education emanating from various sources¹³.

4 Conclusion

DiaCollo serves as an effective automatic tool for the analysis of semantic change with respect to terms and concepts in diachronic perspective. Designed and optimized for the needs of humanities researchers, DiaCollo’s expressive query language and flexibility make it a powerful tool for corpus research. Thorough documentation, tutorials, and references to previous work as well as user-oriented dissemination in the form of workshops and lectures by the CLARIN-D working group “history” make it easy for the inexperienced to learn and provide a useful resource for more experienced users. Actively maintained and supported, DiaCollo continues to evolve and adapt in response to and in co-operation with its user community. Our use cases have shown the necessity of constant shifts between close and distant reading methods, which DiaCollo facilitates. Although ensuring interoperability between tools and maintaining the high standards of data curation necessary for reliable results requires considerable effort across all disciplines, we believe the prospective gain for the scientific community will justify the endeavor.

References

- Evert, S. “Corpora and collocations.” In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, Berlin, Mouton de Gruyter, pp. 1212–1248, 2008.
- Hamp, B., Feldweg, H. “GermaNet – a lexical-semantic net for German.” In Proceedings of the ACL workshop *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.
- Jurish, B. “DiaCollo: On the trail of diachronic collocations.” In K. De Smedt (ed.), *CLARIN Annual Conference 2015* (Wrocław, Poland), 2015.
- Jurish, B. “Diachronic Collocations, Genre, and DiaCollo.” In Whitt, R. J. (ed.), *Diachronic Corpora, Genre, and Language Change*. Amsterdam, John Benjamins, pp. 42–64, 2018.

¹¹ <https://kaskade.dwds.de/~jurish/cac2019/Schule-gb-adj>

¹² <https://kaskade.dwds.de/~jurish/cac2019/ultramontan-freq>

¹³ <https://kaskade.dwds.de/~jurish/cac2019/ultramontan-germanet>

Corpus-Preparation with **WebLicht** for Machine-made Annotations of Examples in Philosophical Texts

Christian Lück

Institut für Neuere deutsche Literatur- und Medienwissenschaft
FernUniversität Hagen, Germany
christian.lueck@fernuni-hagen.de

Abstract

This paper is an outline of an architecture used for harvesting examples from a corpus of philosophical writings. CLARIN-DE's **WebLicht** is used for preprocessing the corpus. The operational mode of a two-stage process on top of it is described. The produced data on example usage in philosophical works is valuable for recent research in literary studies and philosophy.

1 Introduction

Recent research in literary studies and philosophy has underlined the role of examples in the formation of knowledge (Ruchatz, Willer, and Pethes 2007; Schaub 2010; Lück et al. 2013; Güsken et al. 2018–). Until now, however, research on examples has remained in an exemplary mode: Single examples, which are presumed to be central to a discourse, are commented in detail following a more or less hermeneutic method. For research on large amounts of examples, there is simply no data set.

How examples help to learn and sometimes even limit the representation of a topic can only be examined for specific domains of knowledge. Efforts have focused on the domain of the philosophy of aesthetics (Derrida 1978; Schaub 2010; Lück et al. 2013), which only emerged in the 18th century. Aestheticians not only make extensive use of examples but also reflect this usage more frequently than writers in other philosophic disciplines. The DFG-funded research project *Das Beispiel im Wissen der Ästhetik* (Fern-University Hagen) supposes that examples (the rose, the tulip, the Alps, but also cultivated landscapes, etc.) are essential to the foundation of basic terms and distinctions of aesthetics, e. g. the beauty of nature vs. the beauty of art. Aesthetic theorems and doctrines are also closely related to normative notions (“Das Wahre, Schöne, Gute”), and examples are the link between the theoretical and the normative level of discourse. A data set of examples in aesthetics would have an additional value for discourse analysis and the archaeology of knowledge following Foucault (1997). It would enable us to present i) an inventory of examples, to make ii) historical cuts (ger. historische Längsschnitte) that reveal the course of the frequency of certain examples over the researched period—their emergence, boom and disappearance—and may-be to iii) correlate trends in the philosophy of aesthetics with other discourses, e. g. travel literature of the 18th or the colonial discourse of the 19th centuries.

Therefore the project is developing methods for harvesting examples from a corpus that spans roughly 60 philosophical works. It started with manual annotations of examples. But this turned out to be too time-consuming and even very difficult: The manual annotations often span complex phrases, sub-clauses, and sometimes even complete sentences. However, we also noticed that there is (nearly) always a single token which is significant for the example. Most of the times it is a noun, sometimes a full verb or an adjective. And when we studied the term frequencies of a corpus document, we discovered that this token has a low to mid-range frequency. We call this single significant token the *head of an example*. A linguistic foundation of this finding may be given in the information structure (*Thema-Rhema-Gliederung*, Bußmann 1990, pp. 784–786) of a text. More important, we can exploit it in a rule-based process for identifying examples in the corpus. This process will be described after outlining corpus preparation.

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

2 WebLicht for Preprocessing

2.1 Architecture

The WebLicht services, part of CLARIN-DE, is at the very heart of the preparation of the corpus for NLP. (Fig. 1) The works of the corpus come in three different formats and from different digitization providers, which are ranked as follows. 1) If a document is provided by an institution like *Deutsches Textarchiv* (DTA) with high standards of quality management, then this provider is chosen. DTA offers TEI-encoded texts and even an already WebLicht-preprocessed version. Unfortunately, less than a handful of works from our corpus can be found in DTA. 2) But 90% of the corpus documents—even very uncommon ones—are present at Google Books. Since they date from the 18th and 19th centuries, PDF downloads are offered with an excellent OCR-layer, that satisfies the DFG’s demands on digitization quality. But unlike digital-born PDFs scanned old books are a challenge for text extraction. Due to slight variances of the height-offset of the glyphs of a line, tools like *pdftotext* or *PDFMiner* don’t generate acceptable outputs.¹ *LAREX* and other tools from *OCR-D*² must first be trained and then do OCR again. Other methods (e. g. Lungen and Hebborn 2012) are not suitable for scripting so that the results are not easy to reproduce. Therefore a tool called *scannedb-ok* was written. It reconstructs the lines and drops headlines, page numbers, sheet signatures.³ It also tries to repair syllabication, but this is a difficult task due to the frequency distribution of the tokens described by Zipf’s law. 3) *Projekt Gutenberg-DE*, *Zeno.org* or the like are scraped only for a hand full of books which are not present on Google Books apparently due to copyright restrictions. There are very essential works among them, e. g. Immanuel Kant’s *Critique of Judgement*.

The texts are then sent to WebLicht as a Service (WaaS), a REST service, which returns the text segmented into sentences and tokens and returns the lemma and the part-of-speech-tag for each token. The *TCF*-Format returned by WaaS is converted to comma-separated values and then stored in a database. At the heart of the database is a table that represents one token per row (one row for each token of each work in the corpus, with lemma, POS-tag, sentence number, position in a sentence, etc. as columns). Then NLP like outlined in sec. 3 is done using *R*⁴ following the tidy principles of data representation (cf. Wickham and Golemund 2017).

2.2 Reproducibility

There are alternatives to WebLicht: widely known and used NLP-libraries for programming languages like Java, Python or R. What are the pros for WebLicht? With WebLicht one gets state-of-the-art

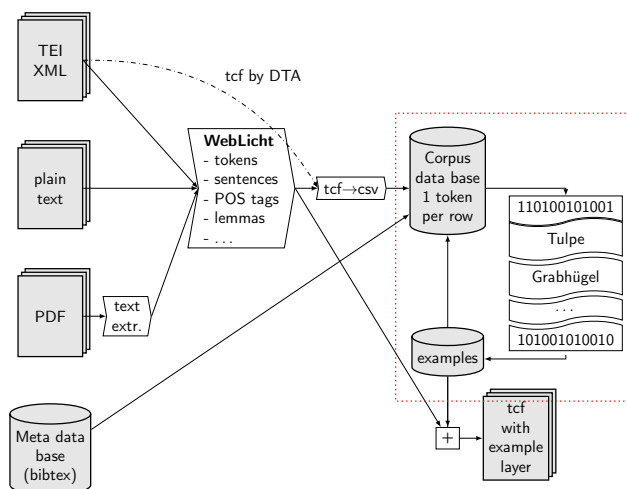


Figure 1: Architecture for text mining used in the project. WebLicht lives in the center of corpus preprocessing. The dotted box demarcates what is done in R. Optionally the tokens and meta data can be transparently kept in a relational database.

¹Cf. <https://www.xpdfreader.com/pdftotext-man.html> and <https://github.com/euske/pdfminer>.

²Cf. <https://github.com/OCR4all/LAREX> and <https://github.com/OCR-D>.

³<http://github.com/lueck/scannedb-ok>

⁴<http://www.r-project.org>

NLP. It requires no time-consuming or error-prone local setup. The preprocessing remains stable when switching between programming languages. WaaS is perfect for scripting (we use Makefiles). All this makes research on top of `WebLicht` more reproducible compared to using a NLP-library of the programming language used in the project. But there are also cons: A library version can be pinned down, whereas there is no notification about upgrades or fixes in `WebLicht`. Is the result stable over time? The community should address the question of reproducibility.

3 Machine-made Annotations of Examples

The process for identifying examples in the corpus is based on the finding, that there is always at least one significant single token within the phrase of the example, which has a low to mid-range frequency. Moreover, most of the times, it is a noun. However, this finding alone does not take us any step further in identifying examples. Zipf's law tells us that there's a vast mass of nouns in the middle- and low-range frequencies. That's why the goal of the algorithm has to be adjusted properly: Do not try to identify all examples, but only those, that appear in a sentence with an example surface marker, like "e. g.". "E. g." marks an example without ambiguity as long as we do not need to expect meta language about example markers. Moreover, this marker can be expected to be in the same sentence as the example.

In the first stage of the process, for each example surface marker the *head of the example*, i. e. the single significant token, is tried to be identified. In the second stage, the resulting list of significant tokens is used to identify examples without surface markers.

3.1 First Stage

In the first stage, a sum of weighted features is calculated for each token of a sentence with an example surface marker. The token with the maximum sum is annotated as the head of the example. The features' weights are picked manually. The features are:

3.1.1 POS-tag

The POS-tag from STTS (Schiller, Teufel, and Stöckert 1999) is mapped to the interval $[0, 1]$ based on manually picked values.

$$f_{POS}(x) := \begin{cases} 1 & \text{if } x \in \{\text{NE, FM}\} \\ 0,8 & \text{if } x \in \{\text{NN}\} \\ 0,5 & \text{if } x \in \{\text{VVINF,} \\ & \text{VVIZU, VVPP}\} \\ 0,4 & \text{if } x \in \{\text{VVF IN}\} \\ 0,2 & \text{if } x \in \{\text{VMINF}\} \\ 0,1 & \text{if } x \in \{\text{VAINF}\} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

3.1.2 Token Frequency

The token frequency is mapped to the interval $[0, 1]$ in such a way, that a hapax legomenon is mapped to 1 and that the value decreases with an increasing frequency. If f_{POS} is 0 for a token, the value should also be 0. Thus, the maximum of the tokens with $f_{POS} > 0$ should be used for normalization. An adaption of the augmented normalized term frequency (Salton and Buckley 1988) is used.

$$f_{tf}(t, D) := \begin{cases} 1 - c \frac{\#(t, D) - 1}{\left(\max_{\{t' | f_{POS}(t') > 0\}} \#(t', D) \right) - 1} & \text{if } f_{POS}(t) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\#(t)$ is the number of times a token occurs in a document and c is a linearity factor with $0 < c < 1$. (The denominator does not equal 0 for natural language documents if at least one non-stop word occurs at least twice.) The lemma frequency f_{lf} is calculated analogously.

3.1.3 Distance from surface marker

The head of an example can be expected to be close to the surface marker. Two types of distance are defined: The distance in tokens f_{dt} is the number of tokens between the marker and the examined token. The distance in commas f_{dc} is the number of commas between the surface marker and the examined token. Let $l(S)$ be the maximum number of tokens before or after the surface marker in sentence S . Let $z(t, s)$ be the number of tokens in sentence S between token t and the surface marker.

$$f_{dt}(t, S) := 1 - \frac{z(t, S)}{l(S)} \quad (3)$$

3.1.4 Result

In Immanuel Kant's *Critique of Judgment*, there are 51 unambiguous example markers "z. B." Results with weights $w_{POS} = 3$, $w_{tf} = 0$, $w_{lf} = 4$, $w_{dt} = 2$, and $w_{dc} = 6$: mihi, Substanz, Bergkristall, Körper, Geister, Rose, Rasenplatzes, Walde, Schönheit, Grabhügeln, Tulpe, Größe, Tiere, Kunstprodukten, Fuß, Affekten, Gebäude, Zorn, Formen, Tulpen, Farben, Lohn, Dichtkunst, Kenntnis, Pferdes, Weib, Genius, Tod, Dichter, Haß, Leuten, Linie, Bau, tun, Parabel, Garten, Zirkels, Eigenschaft, Flüsse, Haus, Körper, Ungeziefer, Winde, Prädikate, Ursache, Made, Wassertiere, Erden, Seele, Ewigkeit, Ewigkeit

3.2 Second stage

As can be seen in the list of example heads in Immanuel Kant's third *Critique*, general concepts are not excluded from examples. The token "Körper", which was correctly identified in one sentence, occurs 33 times throughout the text. The correctly identified token "Größe" occurs 48 times. In both cases, only a few other occurrences, if any, can be interpreted as examples. So there is definitively a need for computing preconditions when using these tokens for the identification of examples without surface markers. Let the document, in which the head has been identified during stage 1, be the source document of a head. Let the rest of the documents in the corpus be the target document. (The source document and the target document may be the same if occurrences in the same document are tested.) Preconditions could be: a) A threshold for the raw token frequency or lemma frequency in the source document: If the threshold is exceeded, occurrences in the target documents do not get annotated. An analogous threshold for the raw frequency in the target document. b) The similarity (e. g. shared lemmas) of the sentence in the source document and the sentence of the target document. c) The similarity not only of the sentences but of an environment of one or several further sentences.

References

- Bußmann, Hadumod. 1990. *Lexikon der Sprachwissenschaft*. 2nd ed. Kröner, Stuttgart.
- Derrida, Jacques. 1978. *La vérité en peinture*. Édition Flammarion, Paris.
- Foucault, Michel. 1997. *Archäologie des Wissens*. Trans. by Ulrich Köppen. 8th ed. Suhrkamp, Frankfurt a. M.
- Güsken, Jessica et al., eds. 2018-. *z. B. Zeitschrift zum Beispiel*.
- Lück, Christian et al., eds. 2013. *Archiv des Beispiels. Vorarbeiten und Überlegungen*. diaphanes, Zürich and Berlin.
- Lüngen, Harald and Mariana Hebborn. 2012. Linguistische Annotation für die Analyse von Gliederungsstrukturen wissenschaftlicher Texte. *Kulturwissenschaften digital. Neue Forschungen und Methoden*. Ed. by Jana Klawitter, Henning Lobin, and Torben Schmidt. Campus, Frankfurt a. M. and New York:155–175.
- Ruchatz, Jens, Stefan Willer, and Nicolas Pethes, eds. 2007. *Das Beispiel. Epistemologie des Exemplarischen*. Kadmos, Berlin.
- Salton, Gerard and Christopher Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management. An International Journal* 24.5:513–523.
- Schaub, Mirjam. 2010. *Das Singuläre und das Exemplarische. Zu Logik und Praxis der Beispiele in Philosophie und Ästhetik*. diaphanes, Zürich and Berlin.
- Schiller, Anne, Simone Teufel, and Christine Stöckert. Aug. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset)*. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart. URL: <http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf> (visited on 01/06/2019).
- Wickham, Hadley and Garret Golemund. 2017. *R for Data Science. Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly, Beijing et al.

Lifespan Change and Style Shift in the Icelandic Gigaword Corpus

Lilja Björk Stefánsdóttir
University of Iceland
Reykjavík, Iceland
lbs11@hi.is

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

Abstract

We demonstrate research on syntactic lifespan change and style shift in Icelandic that is made possible by recent advances in Language Technology infrastructure for Icelandic. Our project extracts data from the Icelandic Gigaword corpus and allows us to shed light on how social meaning shapes the linguistic performance of speakers using big data methods that would not have been feasible for us to use without a corpus of this type.

1 Introduction

In this paper, we describe a case study where we use the recently constructed Icelandic Gigaword Corpus (Risamálheild) (Steingrímsson et al., 2018) in order to examine syntactic lifespan change and style shift in the speech of an Icelandic speaker, former minister of finance, Steingrímur J. Sigfússon. Our study exemplifies how a general-purpose resource for Language Technology can facilitate big data research in the digital humanities if it is carefully curated and made freely available to researchers. Without recent advances in Language Technology infrastructure for Icelandic, our study would have been a prohibitively daunting task, showing that important progress is being made, even in the case of a less-resourced language like Icelandic.

In recent years, studies of language variation and change have increasingly paid attention to linguistic change across the lifespan of an individual. This is interesting because it is widely believed that a critical period for language acquisition constrains the malleability of linguistic abilities (Lenneberg, 1967) and, empirically, the organization of language is indeed rather stable in the adult brain. It is therefore important to improve our understanding of what can change in the language of adults and how. Most current studies on lifespan change have a limited time resolution, typically looking at only 2–3 periods in the speaker’s life (see for example Harrington, 2006, Sankoff and Blondeau 2007, Kwon 2017, MacKenzie 2017, but also Arnaud 1998 and Sankoff 2004). We argue that an improved time resolution is critical for studies of this type as well as a focus on qualitative detail when interpreting quantitative findings.

We examine the variable use of the syntactic process of Stylistic Fronting (SF) throughout the career of an Icelandic politician, Steingrímur J. Sigfússon. We reveal an age grading pattern (e.g., Labov 1994; Wagner 2012) toward less formal usage that is disrupted by a spike in use of the formal variant following the Icelandic economic crash of 2008 when Sigfússon, the leader of the Left-Green Movement, becomes the Minister of Finance and becomes publicly responsible for the fate of the Icelandic economy. We attribute the spike to a dramatic change in his Linguistic Market Value (LMV) in the sense of (Sankoff and Laberge, 1978). This temporary change is then reversed when the left wing government loses its majority in the 2013 election and Steingrímur stops being a Minister in the government as well as the leader of his party. The findings demonstrate how a fine-grained view of syntactic lifespan change yields insights about age-associated usage and status-associated usage as interrelated aspects of the social dimension of language. We also examine the stylistic dimension and suggest that style shift reflects a situational LMV.

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

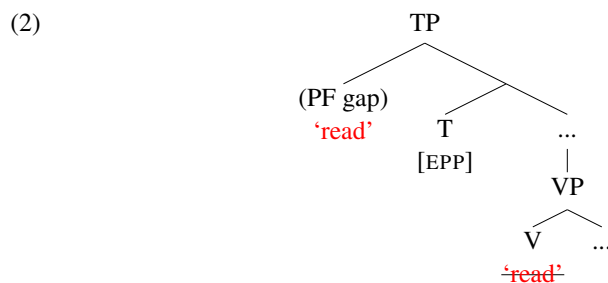
2 The Stylistic Fronting variable

SF is an optional movement in Icelandic of some category to the apparent subject position of a finite clause which does not have an overt subject (Maling, 1980). The envelope of variation includes subject relatives, embedded subject questions and various impersonal constructions. Although the (phrasal) subject position, e.g., Spec,TP, seems to be the target of SF, the moved category is often a head, canonically a **non-finite main verb** that moves in front of a **finite auxiliary** like the passive participle in (1). The second word order variant shown in the curly bracket is a non-SF counterpart. According to one analysis (Holmberg, 2006), SF is one way to satisfy the PF part of the EPP requirement.

The example in (1) demonstrates the variation associated with SF. In a relative clause with a subject gap, the non-finite verb *lesnar* can be moved into the subject gap.

- (1) *Bækur* [_{CP} *sem* {*lesnar eru* / *eru lesnar* } *til skemmtunar*] *eru bestar*.
 books [_{CP} that {**read are** / **are read** } for entertainment] are best
 ‘Books that are read for entertainment are the best ones.’

Structurally, the movement looks as in (2). The T head has some kind of a subject requirement, whose precise formulation is beyond the scope of this paper, but crucially the absence of an overt subject, even in cases where a covert subject is likely to be assumed by many analysts, allows for SF. The PF gap allows any SF-suitable element to move into an apparent subject position, e.g., Spec,TP.



3 Detecting patterns in the Icelandic Gigaword Corpus

The Icelandic Gigaword corpus consists of about 1300 million running words of text and a part of the corpus are parliament speeches. Since the subject of our study is a member of the Icelandic parliament, Alþingi, and due to the fact that Sigfússon is the parliament member who has spoken the most words at the Icelandic parliament (about 6 million words), we make use of the biggest collection of spoken language from the same speaker available to us in order to study the nature of lifespan change and style shift in spoken language. We wrote a Python script that analyzed part of the Gigaword corpus, the parliament speeches given by Sigfússon between 1990-2013, and we extracted sequences with a relative complementizer followed by a finite verb and a non-finite one in either of the two possible word orders. It should be noted that the corpus is not parsed for syntactic structure. Nevertheless, the patterns that we search for are very reliable as confirmed by our manual checks of the extracted data.

To control for various contextual factors (Wood, 2011) we only collected subject relatives with a potential for SF of a non-finite verb. This provided us with 8005 tokens of the SF variable. Each token was coded for SF application, speaker’s age and type of speech (prepared/response).

4 Lifespan change

Use of SF across Sigfússon’s career is shown in Fig. 1. There is a downward trend in the use of SF from age 35 onwards. We interpret this as an age-associated pattern (age grading) resulting from a reduced pressure to conform to the formality demands of the parliament as he gains seniority. However, his role as a leading opposition voice increases his LMV when the economy fails in 2007-2008 and the trend is reversed. When he becomes Minister of Finance (green period in Fig. 1) his SF use rises sharply.

When he stops being a Minister in government in 2013 and steps down as the leader of the Left-Green Movement, returning to being a common opposition MP, his SF returns to pre-economic-crash levels.

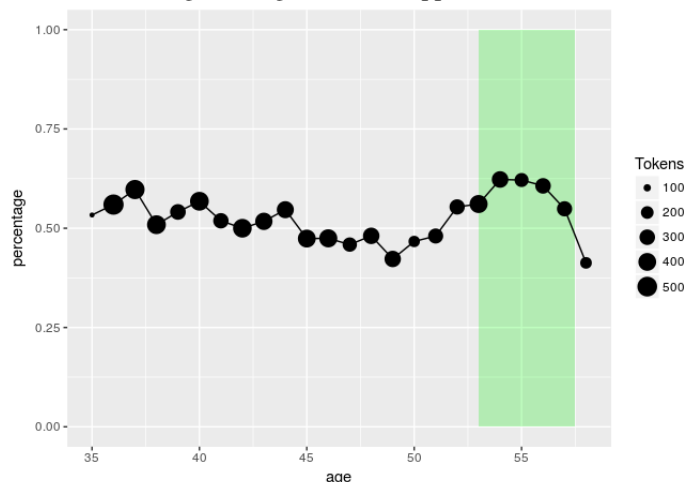


Figure 1: Evolution of SF use in Sigfusson's career.

The main point to be taken from these facts is that various nuances in the development can only be studied in a large digitized corpus. If we had, for example, only the first data point and the last data point, it would not matter how carefully the data were collected and curated; many crucial aspects of the development would simply be missing from the picture.

(3) **Main point about the methodology:**

While community-wide usage evolution is often regular and gradual, individual lifespan change responds rapidly to idiosyncratic sociolinguistic pressures – demanding a high-definition approach.

5 Style shift

In Fig. 2, we split the data between prepared speeches (more SF) and responses (less SF), visualized as a locally weighted regression. Consider now the audience design theory (Bell, 1984) which was constructed to amend some issues with the (Labov, 1972) notion of attention-paid-to-speech. The audience design theory is not the most obvious explanation for a style shift in 8005 tokens given from the same podium in the same room, all of which have the Icelandic parliament as an audience. Of course it is possible to say that preparation simply yields a fundamentally different type of language, but this is nevertheless as close to a fixed audience type as one can imagine.

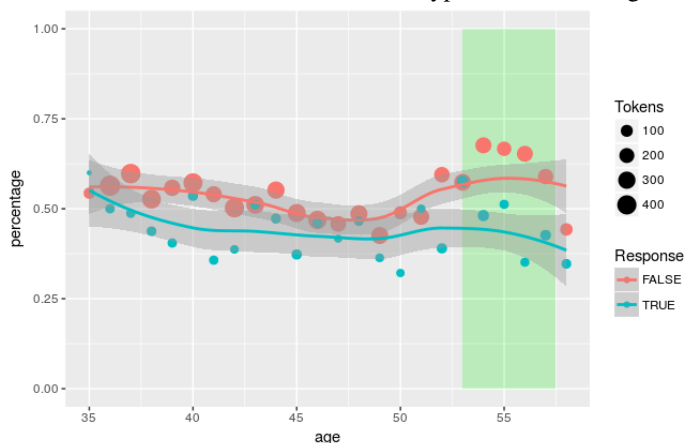


Figure 2: Style shift between prepared speeches (FALSE/red) and responses (TRUE/green)

In the presentation we will discuss our view that usage probability at time p_t is a function of a base probability p_0 , the LMV of the individual and the LMV of the current situation, thus reducing the social and stylistic dimension of variation to two interrelated aspects of the linguistic market. This view can capture LMV via attention-to-speech as well as adaptation to the audience.

$$(4) p_t = p_0 + \text{LMV}(\text{individual}) + \text{LMV}(\text{situation})$$

What we mean by LMV via attention-to-speech is the fact that language which is prepared on paper is likely to be viewed as something that should strive to meet a higher standard with respect to the social evaluation of language. We believe that unifying social properties of the individual and the social properties of the speech situation is a feasible theoretical reduction which is a useful null hypothesis until proven wrong.

6 Summary and future work

Our findings add to much ongoing work on lifespan change and because of the wealth of data that are available in Sigfússon's speeches we get a high definition view of syntactic change across the lifespan. The findings reveal an interplay of age-associated and status-associated factors. In our presentation, we will also discuss how this study demonstrates that qualitative detail is important when interpreting quantitative findings; the computational power that the digital humanities have made available for researchers complements rather than replaces well established methods that focus on attention to detail and context.

References

- Arnaud, Rene. 1998. The development of the progressive in 19th century English: A quantitative survey. *Language Variation and Change*, 10:123–152.
- Bell, Alan. 1984. Language style as audience design. *Language in society*, 13:145–204.
- Harrington, Jonathan. 2006. An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics*, 34:439–457.
- Holmberg, Anders. 2006. Stylistic fronting. *The Blackwell companion to syntax*, pp. 532–565.
- Kwon, Soohyun. 2014. Noam Chomsky's vowel changes across the lifespan. *Selected papers from NAW 42, U. Penn Working Papers in Linguistics*, 20:91–100.
- Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.
- Lenneberg, Eric. 1967. The biological foundations of language. *Hospital Practice*, 2:59–67.
- MacKenzie, Laurel. 2017. Frequency effects over the lifespan: A case study of Attenborough's r's. *Linguistics Vanguard*, 3.1 (2017).
- Maling, Joan. 1980. Inversion in embedded clauses in Modern Icelandic. *Íslenskt mál* 2:175–193.
- Sankoff, Gillian. 2004. Adolescents, young adults and the critical period: Two case studies from 'Seven Up'. *Sociolinguistic Variation: Critical Reflections*, ed. Carmen Fought, 121–139. Oxford University Press, New York.
- Sankoff, Gillian, and Helene Blondeau. 2007. Language change across the lifespan: /r/ in Montréal French. *Language*, 83:560–588.
- Sankoff, David, and Suzanne Laberge. 1978. *The linguistic market and the statistical explanation of variability. Linguistic variation: Models and methods*. Academic Press, New York.
- Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jon Gudnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.
- Wood, Jim. 2011. Stylistic Fronting in spoken Icelandic relatives. *Nordic Journal of Linguistics*, 34:29–60.

Studying Disability Related Terms with Swe-Clarín Resources

Lars Ahrenberg¹, Henrik Danielsson², Staffan Bengtsson³, Hampus Arva¹,
Lotta Holme², Arne Jönsson¹

¹ Department of Computer and Information Science, Linköping University, Sweden
lars.ahrenberg|hampus.arva|arne.jonsson@liu.se

² The Swedish Institute for Disability Research, Linköping University, Sweden
henrik.danielsson|lotta.holme@liu.se

³ The Swedish Institute for Disability Research, Jönköping University, Sweden
staffan.bengtsson@ju.se

Abstract

In Swedish, as in other languages, the words used to refer to disabilities and people with disabilities are manifold. Recommendations as to which terms to use have been changed several times over the last hundred years. In this exploratory paper we have used textual resources provided by Swe-Clarín to study such changes quantitatively. We demonstrate that old and new recommendations co-exist for long periods of time, and that usage sometimes converges.

1 Introduction

Digitisation (with OCR) of textual material previously available only in print has enabled large-scale quantitative studies of the recorded past. Coupled with methodological developments in natural language processing (NLP) and other fields, researchers in the humanities and social sciences can study the past in new and powerful ways. Well known examples can be found in the study of literature (Moretti, 2005; Jockers, 2013), in cultural history (Michel et al., 2011) and language change (Tahmasebi et al., 2018).

It is noteworthy that much of the research so far has been conducted on English data. As the quantity of historical Swedish texts that are digitised is increasing, linguistic change in Swedish, whether "natural" or prompted by technological innovations or by the recommendations of public authorities, can begin to be studied by digital methods. In this study we are concerned with lexical changes in the domain of disabilities. This domain is of special interest in a Swedish setting as the understanding of what disability is, and what it means, has been the subject of much debate, causing new recommendations to be issued from time to time as regards appropriate terminology (see Section 2). To the best of our knowledge this is the first quantitative study of Swedish disability terms.

The study is a collaboration between computational linguists on the one hand and historians and disability researchers on the other. It is ongoing; in this paper we report some early results.

2 The concept of disability (in Sweden)

Several models and perspectives have been discussed and proposed in relation to disability. The traditional way to approach this field has been labelled the *medical* or *individual* model. This is foremost a term that has been introduced by its opponents, as a contrast, and can hardly be said to have its own advocates. The medical model tends to reduce the phenomenon to body functions and bodily deficits. Thus, disability occurs on an individual level, since it is the restrictions caused by physical or mental deviations or flaws that in the end explains why someone experiences problems in everyday life. The inherent logic of the medical model is to a large extent guided by ideas of bodily normality and so much of the attention is directed to compensation.

This way of approaching and understanding disability has been challenged by the environmental turn first materialised in the so-called *social model*, a model that emerged within British disability activism in the 1970s. As opposed to the medical model, disability is rather viewed as the outcome of social, structural and institutional barriers. What turns an impairment into a disability depends on how the society is

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

constructed. If society creates barriers in forms of both physical inaccessibility and degrading attitudes leading to various actions of discrimination the answer is not about normalising measures but to change how society works. According to the social model, disability is about combating all these social barriers.

One objection to the social model has been its alleged neglect of impairments and the body as well as the experience of the individual. As a part of this criticism competing models have been developed. In Scandinavia the *relative*, or *relation model* has gained ground. According to this approach the question of what becomes a disability is not given but comes as a result of the interaction between the individual and the surrounding environment. A person with a certain impairment can be disabled in one specific context or situation but not in another. It all depends on how the environment is constructed and what type of support is available. While the social model's claim of universal barriers, injustice and oppression is difficult to maintain, the relational model is close to it by emphasising that disability must be understood in relation to the environment.

A great breakthrough for the disability movement in Sweden came in the 1970s when the Disability Federation Central Committee introduced a joint disability programme, called A Society for All. As early as the 1960s, the concept of disability in official documents and legal texts included some social model elements, and in the programme A Society for All it was claimed that society and the environment should be designed according to the needs of all citizens. It was not enough to bring the individual to society; society must also be made accessible. An important question for us is to what extent this view of disability is found in official reports and media.

The term *handikapp* ('disability') was introduced as an umbrella term for the many different terms that denoted special types of disability. *Handikappad* ('disabled') was something a person was, but with the introduction of an environment related view, other words such as *funktionsnedsättning* ('functional impairment') and *funktionshinder* ('functional impediment') were recommended. More recently these words, too, have been put into question, and a shift of attention to enabling measures has been proposed signalled by terms such as *delaktighet* ('participation'). These changes are not only replacements of forms but of (desirable) denotations and connotations.

3 Data processing and analysis

Our primary resource for this study are the Official Reports of the Swedish Government (henceforth: SOU¹) from 1922 to 2016 as found at the Swedish Language Bank², the resource hub of Swe-Clarín³.

For the studies on frequencies and word embeddings the texts were lowercased and stop words were removed and grouped into decades. It was necessary to use this coarse granularity as reports covering topics related to disability are unevenly distributed over years.

The SOU-files, especially for the earliest periods, contain many errors due to failing OCR. However, word-based methods are often robust allowing general trends in the data to be captured even in the presence of noise. As it turned out, also word embeddings could be produced, showing plausible relations between terms.

3.1 Frequency changes

An initial list of 60 words referring to disabilities and/or disabled persons over the last 100 years was manually produced by the disability researchers. For lack of space we only report data for the general terms introduced above.

The words are either nouns or adjectives, which means that they occur in Swedish text in a variety of inflectional forms, up to eight for nouns, up to ten for adjectives. They also form derivatives and compounds. We have assumed that sharing of a common stem implies sharing a meaning.⁴ This is a simplification, but does not prevent the discovery of general trends. Thus, we are comparing relative

¹An acronym for Statens Offentliga Utredningar.

²<https://spraakbanken.gu.se/swe/resurs/rd-sou#tabs=information> and <https://spraakbanken.gu.se/swe/resurs/sou#tabs=information>.

³The most recent version at the time of writing being published in July, 2017.

⁴For most words, the stem is identical to the look-up form in a Swedish dictionary. For some words, two forms are required due to stem variation, as in *galen* ('mad' singular) vs. *galna* ('mad' plural), *galning* ('mad person').

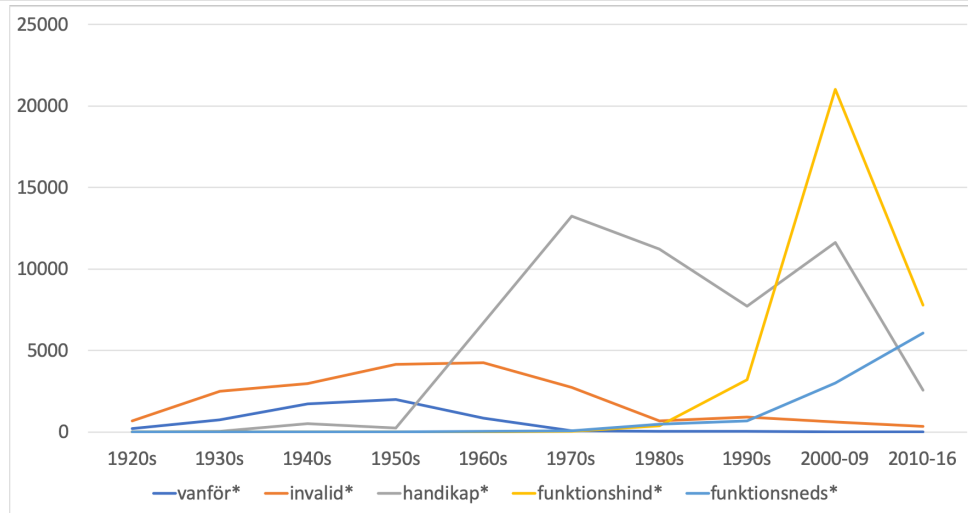


Figure 1: Usage of some Swedish disability terms 1922-2016 by decades.

frequencies within a cohort of terms assumed to cover roughly the same semantic space over a decade. See Figure 1. We can see clear changes of dominant terms since the 1920s:

invaliditet/vanförhet → handikap → funktionsnedsättning/funktionshinder

In agreement with our expectations, these data indicate that there has been a change in the language used to refer to disabilities and persons with disabilities. While the last change is to be expected from the adoption of a relational model, it seems to happen quite slowly and with full force much later than one could expect. Also, the change is not abrupt. Thus, it seems that several disability models are at work simultaneously.

3.2 Terms in context

By looking at the contexts in which a word is used we can gain an understanding of how people use it. We may use a concordancer or simply look at cooccurrences with words in the context. The Korp concordancer⁵, which has a parsed version of the SOU-texts, can display neighbours with different grammatical relations to a word with their frequencies. Korp also enables the generation of concordances for the pair of context word and key word. See Figure 2.

1. svår	233	De som har allvarliga	medicinska handikapp	t ex.
2. psykisk	147		STATENS OFFENTLIGA UTREDNINGAR	(stödjer ej
3. fysisk	148	arbetsföra av 2:a graden innehöll individer med mera avancerade	medicinska handikapp	.
4. grav	63	t enkel och arbetsförmågan kan bedömas med kännedom om det	medicinska handikapp	, som del
5. social	125	a en tillfredsställande arbetsinsats endast om hänsyn tas till deras	medicinska handikapp	vid val a
6. livslång	58	imma betingelser kan dessa som ovan framhållits trots avsevärda	medicinska handikapp	ofta gör
7. motorisk	37	arbetsförhållanden för arbetstagare, som på grund av avsevärda	medicinska handikapp	icke kan
8. oläk	174		Det medicinska handikappets	tidigare
9. neurologisk	27		Det medicinska handikappets	som kar
10. allvarlig	43	hänsyn handikapp i första hand kan det	medicinska handikapp	påverka
11. mental	23	Om en person trots	medicinskt handikapp	bedömt

Figure 2: Adjectival attributes of the noun *handikapp*, 'disability', from 1 billion tokens in the Swedish Language Bank. Apart from the SOU-files all newspaper data and a corpus of novels have been included. The icons to the right of an attribute provide links to a KWIC concordance, where the search word occurs with this particular attribute.

⁵<https://spraakbanken.gu.se/korp>

The figure shows that ‘handikapp’ is often accompanied by attributes referring to extent: *svår*, ‘hard, difficult’, *grav*, ‘grave’, *allvarlig*, ‘serious’, *lätt*, ‘light’, or kind: *psykisk*, ‘mental’, *fysisk*, ‘physical’, *neurologisk*, ‘neurological’, *medfödd*, ‘congenital’, *livslång*, ‘life-long’. There are many overlaps with the corresponding lists for the words *sjukdom*, ‘disease’ and *funktionshinder*, ‘functional impairment’. This clearly gives the impression that the medical model is well represented in the data.

funktionshinder	1970-79	1980-89	1990-99	2000-09	2010-16
1970-79	=	sjukdomstillstånd	sjukdomar	sjukdomar	smärttillstånd
1980-89		=	handikapp	funktionsnedsättning	funktionsnedsättning
1990-99			=	funktionshinder	funktionsnedsättning
2000-09				=	funktionsnedsättning

Table 1: Forward temporal analogies for the term *funktionshinder*.

To obtain richer models we have trained word embeddings for the full corpus of SOU-reports, and for each decade⁶.

Because the training algorithm initializes with random weights, ten models were trained for each decade from the 1970 and forward, resulting in a total of fifty models. To control that the models were stable, the top three words by cosine similarity were checked for each one of the fifty models.

We could see a change in moving from the 1970:ies to the 1980:ies. For the terms *funktionshinder* and *funktionsnedsättning*, the term *handikapp* is one of the three closest neighbours only once in the period 1970-79. In the following period 1980-89, *handikapp* is the closest neighbour for both terms.

We have also compared the vector spaces for different decades using the technique of temporal analogies (Szymanski, 2017). The method enables comparisons of one vector space to another by a global transformation or projection. Each ‘early’ model was paired with each ‘later’ model, e.g. a model from 1970 was paired with all other 1970 models and all models from later decades. After the transformation, cosine similarity was checked again in each model to see if there were any new words showing up. Table 1 shows the closest analogy for the term *funktionshinder*. The picture we got for this word by considering the neighbours in each decade, is confirmed. In the 1970:ies this term was used differently, analogously to words such as *sjukdomar*, ‘diseases’ in later years. From 1980 onwards it shows more affinity with the terms *handikapp* and *funktionsnedsättning/funktionsnedsättningar*.

4 Conclusions

Quantitative word-based investigations of even fairly noisy textual data may corroborate and enrich qualitative approaches and reveal patterns and trends in language use that can be compared to advice and recommendations. Our analysis of the use of Swedish disability terms in resources made available by Swe-Clarín partners indicates that, while recommendations have effects, they are delayed, and that several frames of thinking about disability live along side by side.

References

- Matthew L. Jockers. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
- Franco Moretti. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
- Terrence Szymanski. 2017. Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. In *Proceedings of the 55th Annual Meeting of the ACL (Short papers)*, pages 448–453.
- Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to diachronic conceptual change. *CoRR*, abs/1811.06278.

⁶We used the Word2Vec CBOW model as implemented in GenSim (<https://radimrehurek.com/gensim/index.html>)

To Ask or not to Ask: Informed Consent to Participate and Using Data in the Public Interest

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

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

Alexandros Nousias
Nousias/Linardos Business &
Legal Consultants
alexandros.nousias@
gmail.com

Abstract

The development and use of language resources often involve the processing of personal data. Processing has to have a legal basis. The General Data Protection Regulation (GDPR) provides several legal grounds. In the context of scientific research, consent and public interest are relevant. The main question is when researchers should rely on consent and when on public interest to conduct research. Both grounds have their advantages and challenges. For comparing them, the Clinical Trial Regulation is used as an example.

1 Introduction

Language resources (LRs) contain material subject to various legal regimes. For instance, they may contain copyright protected works, objects of related rights (performances) and personal data. This affects the way language resources are collected and used. Intellectual property issues relating to language resources have previously been addressed by Kelli et al. (2015). The general approach to dealing with personal data in research is outlined by Kelli et al. (2018), where they discuss how processing personal data without consent as the legal basis is possible in various EU countries.

Personal data issues are relevant for language resources, given that they potentially contain oral speech or written text, which relate to a natural person. The GDPR¹ (General Data Protection Regulation, 2016) provides a general framework for personal data protection. This article primarily outlines the regulatory framework for processing personal data for research purposes.

According to the GDPR, there are several legal grounds for processing personal data. We focus on two legal grounds relevant for research, i.e. consenting to use personal data and using data for research in the public interest. When using data for research purposes in the public interest, we also need to consider whether it is feasible to inform the data subjects.

If we get data directly from the data subjects, the GDPR requires that the data subjects are informed (Art. 13), i.e. we need to get confirmation from the data subjects that they are aware of our activity. If on the other hand, we reuse personal data from large databases or publicly available sources, it is not always possible to inform the data subjects.

However, the focus of this study is to show how one can collect personal data directly from the data subjects while still using research in the public interest as the legal ground. We call this model confirmation to participate in research of public interest, and show that this is not just a hypothetical model but it is already in use in the CTR (Clinical Trial Regulation, 2014). Framed in the terminology of the GDPR, the CTR model² is based on processing data for a task carried out in the public interest. The data subject nevertheless has to be informed about the processing and must consent to participating in the

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

¹ The GDPR is applicable in all EU Member States from 25 May 2018. It replaces the Data Protection Directive (1995).

² https://edpb.europa.eu/our-work-tools/our-documents/avis-art-70/opinion-32019-concerning-questions-and-answers-interplay_en

trial, i.e., the data subject gives informed consent to participate in the research and may end participation at any time. However, according to the CTR, the data subject does not give specific consent to process personal data in the way in which consent is defined in the GDPR. The CTR consent has therefore also been labeled “broad consent”³, “ethical consent” or “consent to participate”, which has important consequences for the right of the data subject to limit processing of the data (e.g., the right to withdraw GDPR consent at any time).

Other types of research than clinical trials can also be carried out in the public interest using the same model for processing personal data, i.e., by *confirmation to participate in research carried out in the public interest*. Since key concepts of the data protection framework (personal data, data subject, etc.) are addressed in previous CLARIN publications, they are not repeated here.

In this abstract, we only explore the consequences of GDPR enabling research in the public interest as a legal ground for doing scientific research vs. using GDPR consent as the legal ground. In the full paper, we will also outline the documents that are needed to implement both of them in practice.

2 Processing personal data for research purposes according to the GDPR

Processing of personal data for research purposes can be illustrated with the graph in Figure 1.

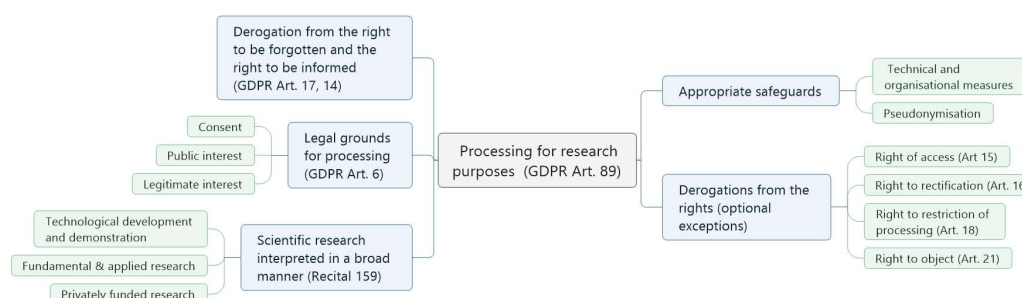


Figure 1. Processing of personal data

The GDPR provides six legal grounds for processing personal data: 1) consent; 2) performance of a contract; 3) compliance with a legal obligation; 4) protection of the vital interests; 5) the public interest or in the exercise of official authority; 6) legitimate interest (Art. 6). The processing for research purposes is not an individual legal ground. The processing for research purposes can rely on consent or the performance of a task carried out in the public interest⁴.

The processing based on the data subject’s consent offers the highest possible protection for the data subject through the following mechanism:

- 1) the consent has to be freely given, specific, informed and unambiguous (Art. 4 (11));
- 2) the data subject can withdraw the consent without any detriment (Art. 7 (3));
- 3) the burden of proof lies with the controller (Art. 7 (1))

WP29⁵ (Article 29 Working Party, 2014: 13) explains that consent “focuses on the self-determination of the data subject as a ground for legitimacy. All other grounds, in contrast, allow processing – subject to safeguards and measures – in situations where, irrespective of consent, it is appropriate and necessary

³ Chassang (2017) discusses the difference between the “broad consent“ in CTR and the specific consent in GDPR (as well as in the preceding directive from 1994) where he regards CTR as *lex specialis*. However, CTR predates GDPR, which now accommodates the CTR-style consent under research in the public interest allowing us to apply it also to other areas of research with public interest as the legal ground.

⁴ Note that legitimate interest as a legal ground for research needs to be argued when one cannot claim to be acting with permission or in the interest of the data subject (legal grounds 1, 2 or 4 in Art. 6) or in the interest of the state (legal grounds 3 or 5 in Art. 6).

⁵ The Article 29 Working Party (WP29) was the independent European working party that dealt with issues relating to the protection of privacy and personal data until 25 May 2018, at which point it was succeeded by the European Data Protection Board.

to process the data within a certain context in pursuit of a specific legitimate interest⁶. In case the acquisition of consent is complicated or administratively burdensome (e.g., anonymous web posts, legacy resources, public videos and so forth), some other legal ground than consent is clearly needed.

According to WP29, the performance of a task carried out in the public interest is also a ground for processing personal data in the research context (Article 29 Working Party, 2014: 21-23). The concept of research in the public interest⁶ can usually be invoked by research projects affiliated with universities or research institutions having a legal mandate to do research in the public interest⁷, but it also allows for companies acting in the public interest on behalf of a Member State, e.g., to ascertain the safety and/or efficacy of a procedure performed in addition to normal clinical practice as outlined in the CTR.

Before addressing specific requirements concerning the processing of personal data for research, it is necessary to outline the concept of research in the data protection context. The GDPR defines research broadly so that it covers “technological development and demonstration, fundamental research, applied research and privately funded research” (Recital 159). The GDPR provides the following requirements for processing data for research purposes (Art. 89):

- 1) processing for research purposes is subject to appropriate safeguards. The safeguards ensure that technical and organisational measures are in place in particular to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner;
- 2) the Member States may limit the following data subject’s rights for research purposes (optional limitations):
 - a) the right of access by the data subject (Art. 15);
 - b) the right to rectification (Art. 16);
 - c) the right to the restriction of processing (Art. 18);
 - d) the right to object (Art. 21);

The implementation of the optional limitations varies by country and they are exemplified and discussed in Kelli et al (2018). However, there are two mandatory limitations⁸ to the rights of data subjects with regard to research data: 1) the right to be forgotten⁹ and 2) the right to be informed about the processing.¹⁰

3 Data Processing according to the Clinical Trial Regulation

The Clinical Trial Regulation (CTR) entered into force on 16 June 2014, but the timing of its application depends on the development of a fully functional EU clinical trials portal and database, which will be confirmed by an independent audit. The Regulation becomes applicable six months after the European Commission publishes a notice of this confirmation. The entry into application of the Regulation is currently estimated to occur in 2019. The GDPR makes a reference to CTR for special requirements for consent to the participation in scientific research activities in clinical trials (Recital 161).

In CTR (Art. 2 (2) 21), *‘informed consent’ means a subject’s free and voluntary expression of his or her willingness to participate in a particular clinical trial, after having been informed of all aspects of the clinical trial that are relevant to the subject’s decision to participate or, in case of minors and of incapacitated subjects, an authorisation or agreement from their legally designated representative to include them in the clinical trial.* This informed consent constitutes a confirmation to participate in a clinical trial. Special requirements for the informed consent are provided in Chapter V of the CTR.

About the withdrawal of informed consent, CTR says (Recital 76) *‘... while safeguarding the robustness and reliability of data from clinical trials used for scientific purposes and the safety of subjects participating in clinical trials, it is appropriate to provide that, without prejudice to Directive 95/46/EC [now replaced by GDPR], the withdrawal of informed consent should not affect the results of*

⁶ According to the GDPR, processing is lawful if it is “processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller” (Art. 6 e).

⁷ For instance, according to the Estonian Organisation of Research and Development Act (1997) a research and development institution is a legal person or an institution in the case of which the principal activity is carrying out basic research, applied research or development, or several of the aforementioned activities (§ 3 (1) clause 1).

⁸ Mandatory limitations are directly applicable. They do not need to be incorporated into the national laws.

⁹ GDPR (Art. 17 (3) d.)

¹⁰ GDPR (Art. 14 (5) b.)

activities already carried out, such as the storage and use of data obtained on the basis of informed consent before withdrawal. This means that the withdrawal of informed consent only implies that the data subject stops participating in the trial. Data collected during the participation can still be stored, e.g. for verifying the results of activities already carried out.¹¹

The legal basis for a clinical trial is research on behalf of a Member State to ascertain the safety and efficacy of a procedure performed in addition to normal clinical practice, which is the prototypical case for research in the public interest. However, for clinical trials, the approval of an ethics committee is also needed, because a non-standard clinical procedure will be applied potentially affecting the well-being of the human subjects.

In GDPR terminology, the CTR can be restated as the data subject's informed consent to participate in research in the public interest.¹²

4 Discussion of consequences for other types of research

Other types of research than clinical trials can also be carried out in the public interest by research projects affiliated with institutions having a legal mandate to do research in the public interest using the same model for processing personal data. If research is carried out in the public interest and data is reused from large databases or public sources, it is not always possible to inform all data subjects about the processing, in which case a public record of processing activities is deemed sufficient.

However, if data is collected directly from the data subjects, and the legal basis according to GDPR is *research in the public interest*, it may be useful to avoid confusion among data subjects and researchers alike by naming the verification for having provided mandatory information "*confirmation to participate in research carried out in the public interest*". The confirmation relies on informed consent to participate, so the consent is given only with regard to the participation and not with regard to the processing of the personal data, for which the legal basis is research in the public interest.

Since research is conducted in the public interest, the data subject's right to be forgotten is mandatorily limited by GDPR (Art. 17 (3) d) to the extent that processing is necessary for research purposes in so far as the right to be forgotten is likely to render impossible or seriously impair the achievement of the objectives of that processing. In many cases, the right to be forgotten could impair the replicability of the research.

It should be noted that data collected in this way can only be used for research purposes. According to the GDPR¹³, such data can still be reused for other research purposes GDPR (Art. 5), but the transfer of the data to another research project needs to be protected to make sure that the data is processed only for research purposes¹⁴. If one wishes to make such personal data publicly available, e.g. as an illustrating example or a video clip potentially identifying the data subject on the internet or at a conference, GDPR consent must be acquired.

The legal framework for processing personal data for research purposes is based on the GDPR and national laws of the EU Member States. This means that in addition to the mandatory limitations of the rights of data subjects enforced by the GDPR, researchers that wish to develop language resources containing personal data may have further rights to maintain the research data integrity through nationally implemented limitations to the rights of data subjects.

5 Conclusion

The development and use of language resources often involve the processing of personal data. To process data for research purposes according to the GDPR, it is possible to invoke research in the public interest as the legal basis for publicly funded research projects carried out at research institutions with a

¹¹ The GDPR has a similar approach. It provides that "The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal" (Art. 7 (3)). However, GDPR also provides for storage extension for research (Art. 5) and a research purpose extension (Art. 5).

¹² As data is collected for research purposes, there is also a limited right to be forgotten (Art. 17).

¹³ GDPR is applicable only within the EU, but many countries have agreements with the EU, see e.g. <https://gdpr-info.eu/issues/third-countries/>

¹⁴ The reuse still needs to answer to general GDPR requirements such as the data minimisation principle. According to (Art. 89(1)) and (Recital 156), for further processing, the controller should also assess the feasibility to fulfil the reuse purposes by processing anonymous or pseudonymous data.

legal mandate to do research in the public interest. If language data is collected directly from data subjects, they have to be informed about the data processing so that they can opt-in by confirming their willingness to participate. If they no longer wish to participate, they have a right to stop, but they do not automatically have a right to be forgotten and their personal data may be reused for other research purposes. This model for using and reusing personal data is already established in the CTR, but it can be extended to other domains as well through the provisions of the GDPR.

In the full paper, we will also discuss how to implement this in practice through a template for a Privacy Notice for Scientific Research for informing the data subjects when collecting data. We will also discuss the key points of a Data Transfer Agreement for personal data that can be added as an appendix to the CLARIN RES licensing templates.

Acknowledgements

We would like to thank the Centre of Estonian Language Resources, FIN-CLARIN and CLARIN-EL for their support.

References

- Article 29 Working Party. 2014. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Adopted on 9 April 2014. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (29.3.2018)
- Chassang G. 2017. The impact of the EU general data protection regulation on scientific research. *Ecancermedicalscience*, 11, 709. doi:10.3332/ecancer.2017.709. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5243137/> (21.8.2019)
- Clinical Trial Regulation. 2014. Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use. Available at: <https://ec.europa.eu/health/human-use/clinical-trials/regulation> (31.1.2019)
- Data Protection Directive. 1995. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995 p. 0031 – 0050. Available at <http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&qid=1522340616101&from=EN> (29.3.2018)
- General Data Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. OJ L 119, 4.5.2016, p. 1-88. Available at <http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679> (29.3.2018)
- Aleksei Kelli, Kadri Vider, and Krister Lindén. 2015. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13–24. Available at <http://www.ep.liu.se/ecp/article.asp?issue=123&article=002> (28.3.2018)
- Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Penny Labropoulou, and Pavel Straňák. 2018. Processing personal data without the consent of the data subject for the development and use of language resources. Selected Papers from the CLARIN Annual Conference 2018, October, 2018, Pisa, Italy. Linköping University Electronic Press, Linköpings universitet. Available at: <http://www.ep.liu.se/ecp/article.asp?issue=159&article=008> (28.8.2019)
- Organisation of Research and Development Act. 1997. Entry into force 2.05.1997. English translation available at <https://www.riigiteataja.ee/en/eli/513042015012/consolide> (21.1.2019)

Data Collection for Learner Corpus of Latvian: Copyright and Personal Data Protection

Inga Kaija

Institute of Mathematics and
Computer Science,
University of Latvia;
Riga Stradiņš University, Latvia
inga.kaija@rsu.lv

Ilze Auziņa

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

Abstract

Copyright and personal data protection are two of the most important legal aspects of collecting data for a learner corpus. The paper explains the challenges in data collection for the learner corpus of Latvian “LaVA” and describes the procedure undertaken to ensure protection of the texts’ authors’ rights. An agreement / metadata questionnaire form was created to inform the authors of the ways their texts are used and to receive the authors’ permission to use them in the stated way. The information, permission, and the metadata questionnaire are printed on one side of an A4 size paper sheet, and the author is supposed to write the text on the other side by hand, thus eliminating the need to identify the author of the text separately. After scanning and adding to the corpus, the text originals are returned to authors.

1 Introduction

Learner corpora have become increasingly popular, and the demand for such corpora to become available to a wider scope of researchers is growing. However, the creation of publicly available learner corpora includes dealing with personal data protection and copyright issues. A learner corpus of Latvian “LaVA” (Latvian Council of Science Grant Development of Learner corpus of Latvian: methods, tools and applications. No. lzp-2018/1-0527) is being created and it will be publicly accessible, so these legal issues have to be addressed while still enabling researchers to collect relevant metadata about possible factors impacting language learning outcomes.

There have been efforts to create templates for contracts to help deal with the copyright issues when collecting data for research¹. While they can be extremely helpful, in the case of creating a learner corpus a more specific compact document is useful where the exact aims and rules of using the texts are described.

In order to protect learners’ rights when collecting their texts, an agreement / metadata questionnaire and the procedure of text collection was developed. The present paper lists the main legal and ethical principles considered and describes how the data collection process is carried out.

2 Regulations

The regulations regarding personal data protection and copyright issues that concern learner corpus creation in Latvia have been previously described in comparison with the relevant regulations in Lithuania (Znotiņa, 2016). We further list the main legal documents and principles to be observed in each of those areas.

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

¹ For example, see <http://www.meta-net.eu/meta-share/licenses>

2.1 Copyright

The main document regulating copyright protection in Republic of Latvia is the Copyright Law (AL, 2000), and it states that:

- texts written as a part of study process are protected by copyright, unless otherwise stated in the study agreement between the author and the study institution;
- in order to make the text (or part of it) available to the public, a written permission must be received from the author;
- the author has the right to decide to be recognized as an author and to decide when, how many times etc. the work can be accessed.

In order to comply with the regulations, a standardized form for all authors of the texts in corpus must be created.

2.2 Personal data protection

Protection of personal data in Republic of Latvia is regulated by the Personal Data Processing Law (FPDAL, 2018) as well as one of the most influential regulations regarding personal data protection in European Union, the GDPR (2016). Both of them emphasize the ability to identify a person as a criterion for defining personal data.

The corpus “LaVA” is a beginner learner corpus, and the topics offered for writing to the beginner students often inherently include telling one’s own or other people’s data (e.g. “Me and my family”). Thus, it is of utmost importance to eliminate the possibility that such personal data would be made publicly available.

3 Data collection for the learner corpus of Latvian

The main principles of the agreement / questionnaire form are the same ones already used in the learner corpus of the second Baltic language “Esam”² (Znotiņa, 2018), but data collection is carried out in a different way. In “Esam”, the permissions to use the data were acquired long after the texts were written (in some cases, several years), and all texts were additionally anonymized. In the case of “LaVA”, the learners know the texts are going to be included in the corpus when they write them. Besides, the texts in “LaVA” are not further anonymized by the project team, and the data is collected by various people, so the procedure is regulated more strictly in order to maintain uniformity in the received data and information given to the authors.

3.1 Contents of the agreement / questionnaire

An agreement / questionnaire form was created for data collection of the corpus “LaVA”. It is written in English because English is used as an intermediary language in studies of Latvian, so all authors speak this language well. The form is offered to all authors of the texts expected to be included into the corpus, and every text is only included into the corpus after a signed copy of the form is received from the author.

The form is printed on one side of an A4 size paper sheet (for layout, see Picture 1) and includes three parts – an information letter, a permission, and a metadata collection questionnaire (information about the author). The former consists of:

- basic information about the project, the institutions that are carrying it out, and contact information;
- brief instructions for learner;
- information about the security of data on the server used for the corpus and privacy;
- explanation on expressing one’s will regarding participation in the project (i.e. what to do if the author decides they no longer want their texts to be used in the corpus).



² Available online: <http://www.esamkorpuss.lv>

The permission includes seven statements the author agrees to comply with by signing the form:

- The author agrees that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus.
- The author confirms that none of the data in this text can lead to identification of any existing people.
- The author agrees that the text is anonymous and their name is not mentioned anywhere on the corpus website or its public documentation. Each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author.
- The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms.
- The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched an unlimited amount of times.
- All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.).
- The author will have the right to withdraw their consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. The author is aware of this opportunity as a data provider.

Finally, the metadata collection questionnaire requests the author to provide some information about factors that may influence their target language production: age, gender, mother tongue(-s), other spoken languages, the length of residence in Latvia, and the number of semesters studying Latvian language in a higher education institution.

The date, signature, name, and surname of the author is needed to ensure the author's full agreement with the aforementioned statements in the permission, but is not included in the metadata of the corpus.

<p>Information letter of the project researcher group for Latvian learners</p> <p>Dear student, The project <i>Development of Learner Corpus of Latvian: methods, tools and applications</i> (Project No. lzp-2018/1-0527) is being implemented at the Institute of Mathematics and Computer Science, University of Latvia (IMCS UL) since September 2018. The goal of the project is to create an error-annotated Latvian language learner corpus and develop corpus-based teaching materials. The project is financed by Latvian Council of Science; the project leader is senior researcher of IMCS UL Dr. philol. Ilze Auziņa (e-mail: ilze.auzina@lumii.lv).</p> <p>What do you have to do? Please read carefully and sign the Permission that you agree to allow the text written during your Latvian language studies to be included in the Latvian learner corpus. Complete the questionnaire and provide the necessary information for the further use of the text in research. On the other side of the page, write an essay on the topic that the lecturer has assigned to you.</p> <p>Data storage and privacy Collected data will be stored at the IMCS UL on the password protected server. The data stored will be completely anonymous. A unique identifier will be assigned to each data provider. After the end of the project <i>Learner Corpus of Latvian</i> will be publicly available on the corpora website of IMCS UL.</p> <p>Participation Participation is voluntary. Over the course of the project, you may request that texts written by you are removed from the database and refuse to participate without specifying the reason. This should be done by informing the group of researchers. In case of refusal, all materials collected will be deleted.</p> <p>On behalf of the project team of researchers, <i>Ilze Auziņa</i>, IMCS UL senior researcher</p> <p> Institute of Mathematics and Computer Science University of Latvia</p> <p> Latvijas Zinātnes padome</p>	<p style="text-align: center;">PERMISSION</p> <p>I agree that this text, written in 2019, can be included in the <i>Learner Corpus of Latvian</i> and, as a part of the corpus, can be made publicly available in various forms, fully or partly, with such conditions:</p> <ul style="list-style-type: none">• I agree that the corpus is available for free and is made for scientific and teaching purposes. The authors do not receive any financial reward for having their texts included in the corpus.• I confirm that none of the data in this text can lead to identification of any existing people.• I agree that the text is anonymous and my name is not mentioned anywhere on the corpus website or its public documentation. Each author receives an anonymous code which makes it possible to recognize several texts written by the same author but does not reveal the identity of the author.• The data included in the corpus can be cited in the educational materials, research papers, and other work in various forms.• The corpus and all materials included in it can be publicly accessible for an unlimited period and can be viewed and researched unlimited amount of times.• All texts included in the corpus can have linguistic information added to them (e.g. error corrections, part-of-speech annotation, etc.).• I will have the right to withdraw my consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. I am aware of this opportunity as a data provider. <p style="text-align: center;">INFORMATION ABOUT THE AUTHOR</p> <p>Age: _____ Gender: _____ Mother tongue (-s): _____ Other languages you speak: _____ How long have you been living in Latvia? _____ For how many semesters have you been learning Latvian language? <input type="checkbox"/> This is the first semester. <input type="checkbox"/> This is the second semester. <input type="checkbox"/> Other (please specify): _____</p> <p>_____ Data Signature Name, surname</p> <p style="text-align: center;">THANK YOU!</p>
--	--

Picture 1: The layout of the agreement / questionnaire form

3.2 Data collection procedure

The authors of the texts are all higher education students who have been living in Latvia for a relatively short time and are learning Latvian language at the beginner level for the first or second semester. Teachers are allowed to choose the desired topic and length of the text, and study materials can be used when writing. The teachers who collect the texts instruct the students about the copyright and personal data protection system used in the project, and remind them that regardless of the topic no real personal information should be included in the text. If the topic contradicts this idea (e.g. “My friends and my family”), students are instructed to write about imaginary people or replace the real information with false one.

After the texts are digitized for inclusion in the corpus, the originals are given back to the teacher who corrects them according to the needs of the pedagogical process, and then hands the texts back to the students, once more reminding them about the possibility to revoke the permission if need be (such as accidental inclusion of real personal data etc.).

4 Conclusions

The agreement / metadata collection questionnaire form used in the learner corpus “LaVA” helps minimise the amount of additional paperwork involved in the creation of the corpus and gives learners a chance to exercise their rights.

The form can be used as a basis for agreements in data collection for other learner corpora in countries which have similar personal data and copyright protection regulations.

References

- Autortiesību likums*, 48/150 (2059/2061), 27.04.2000. [Viewed on April 29, 2019]. Available online: <https://likumi.lv/doc.php?id=5138>
- Fizisko personu datu apstrādes likums*, 132 (6218), 04.07.2018. [Viewed on April 29, 2019]. Available online: <https://likumi.lv/ta/id/300099-fizisko-personu-datu-apstrades-likums>
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, OJ 2016 L 119/1.
- Inga Znotiņa. 2016. Valodas apguvēju korpuss Latvijā un Lietuvā: autortiesības un personas datu aizsardzība. *Vārds un tā pētīšanas aspekti*, 20(2):219–227.
- Inga Znotiņa. 2018. *Otrās baltu valodas apguvēju korpuss: izveides metodoloģija un lietojuma iespējas*. Doctoral dissertation. Liepāja University, Liepāja.

Liability of CLARIN Centres as Service Providers: What Changes with the New Directive on Copyright in the Digital Single Market?

Pawel Kamocki
IDS Mannheim
pawel.kamocki@g
mail.com

Erik Ketzan
Birkbeck,
University of
London
eketza01@mail.b
bk.ac.uk

Julia Wildgans
IDS Mannheim /
Universität
Mannheim
j.wildgans@ggoog
lemail.com

Andreas Witt
IDS Mannheim /
Universität
Mannheim /
Universität
Heidelberg
witt@ids-
mannheim.de

Abstract

Providing online repositories for language resources is one of the main activities of CLARIN centres. The legal framework regarding liability of Service Providers for content uploaded by the service users has very recently been modified by the new Directive on Copyright in the Digital Single Market. A new category of Service Providers — Online Content-Sharing Service Providers (OCSSPs) — is subject to a complex and strict framework, including the requirement to obtain licenses from rightholders. The proposed paper discusses these recent developments and aims to initiate a debate on how CLARIN repositories should be organised to fit within this new framework.

1. Introduction

One of the main activities of CLARIN centres is to provide online services for their users, such as online repositories. However, the content uploaded by users of such services can sometimes be of infringing nature, and researchers are (or should be) aware that language resources can violate many rules, from copyright through related rights (such as the *sui generis* database right) to data protection.

The question of liability for hosting content (such as language resources) uploaded by users of scientific repositories has not attracted the attention that it deserves, although some work on this topic has been presented at conferences (Kamocki, 2014). One might think that someone who merely provides an online service (e.g. stores data) should not be liable for the illegal activities of its users. Under the current rules, this statement is largely true, and this common-sense point of view has been reflected in the normative framework for almost the past twenty years (i.e. from the beginning of the participative Web).

By pure law, however, service providers may be found liable for prejudice caused by users. Liability requires three fundamental elements: breach, prejudice and a causal link between the two (causation). If there is a breach of law (e.g. copyright infringement or unlawful processing of personal data) that causes prejudice, this prejudice can be causally linked to the actions of a service provider thereby making him liable. For example, if sensitive information related to a person's health or sexual orientation is communicated to millions of Internet users via an Internet service (e.g. social media), the prejudice suffered by the victim is in fact directly caused by the social media website which made this information available to its users.

This does not mean that the user is not liable for his actions — he or she can also be sued for damages, but from the victim's perspective, the service provider is often more in the position to pay substantial damages. Not only is the service provider easier to identify, but also, as a company or an insti-

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

tution, it is expected to be more solvent than an individual user, and possibly also more inclined to settle to avoid damage to its reputation.

In order to promote the development of online services, both the United States (see the Digital Millennium Copyright Act 1998) and the European Union (see below) adopted special rules (called liability exemptions) to protect service providers from liability for illegal actions of their users, thus creating a mechanism described as *Safe Harbor*. In a world without the Safe Harbor, services like Facebook, Twitter or Youtube would not have been able to thrive due to the massive potential liability costs involved.

Nowadays, however, it seems that things are slowly beginning to change, especially in Europe, where many users feel that they live in a world dominated by huge, seemingly omnipotent service providers such as Google, Amazon or Facebook. Apparently, European legislators also think that some large and ‘powerful’ service providers not only do not deserve preferential treatment, but should be more or less openly discriminated against, to favour the development of local businesses¹. In response to such sentiments, the first big step in new provisions to protect user interests was made by the recently adopted Directive on Copyright in the Digital Single Market (see below).

Below, we discuss whether and how this situation will affect smaller service providers such as CLARIN centres.

2. Liability of Service Providers from e-Commerce to the Digital Single Market

Under EU law, the liability of service providers was harmonised in the Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (commonly referred to as the e-Commerce Directive). In 2019, this framework has been modified by the Directive on Copyright in the Digital Single Market.

‘Service provider’ is defined as any natural or legal person providing an information society service. ‘Information society service’ is further defined as “any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services” (art. 1, Directive 2015/1535). It is important to note, however, that according to this definition the service does not have to be provided for remuneration to qualify as ‘information society service’. Therefore, hosting providers who host data for free (like Facebook, which hosts users’ accounts without any remuneration) are still covered by the discussed framework.

2.1. Liability of Service Providers in the e-Commerce Directive

The e-Commerce Directive concerns three types of information society services: mere conduit (art. 12; defined as transmission of information in or provision of access to a network), caching (art. 13; defined as intermediate and temporary storage of information, performed for the sole purpose of making its onward transmission more efficient) and hosting (art. 14). This last category is the most relevant for CLARIN Centres.

Hosting is defined as storage of information provided by the user. The hosting provider is not liable for the stored information on condition that he does not have actual knowledge of its content and, upon obtaining such knowledge, acts expeditiously to remove or to disable access to the information (hence the commonly used *notice-and-take-down* procedures). Article 15 further states that service providers have no general obligation to monitor the content that they store or transmit, or to actively seek facts or circumstances indicating illegal activity.

In the participative Web, drawing a line between the activities of the user (content provider) and those of the service provider is not always an easy and straightforward task. However, the distinction (content provider vs. service provider) is crucial from the legal standpoint.

When the e-Commerce Directive was adopted (2000), hosting providers were typically merely offering storage space, but now, they often play a much more active role in presenting content — it is quite impossible for a visitor not to know that this particular piece of content is hosted by Facebook, Youtube or Amazon. Quite the contrary, it is possible for an ordinary user to overlook the identity of the content provider, but paradoxically it’s usually the service provider who attracts viewers. In other

¹ Cf. especially the ‘GAFAs tax’ initiative launched by the French government: <https://www.gouvernement.fr/en/gafa-tax-a-major-step-towards-a-fairer-and-more-efficient-tax-system> (accessed 26.08.2019).

words, with the participative Web, the distinction between the content providers and the service providers gets blurred. It is not surprising, therefore, that the liability of the latter is often sought.

For the Court of Justice of the European Union (CJEU, joined Cases C-236/08 to C-238/08, 23 March 2010), in order to qualify for the liability exemption, the Service Provider must meet the criteria set forth in recital 42 of the e-Commerce Directive, according to which its activity must be “of a mere technical, automatic and passive nature, which implies that the information society service provider has neither knowledge of nor control over the information which is transmitted or stored”. The Court applied the exemption e.g. to Google (regarding the GoogleAds service), but a new ruling is about to come: last year, the German Federal Supreme Court referred a case about YouTube’s liability to the Court of Justice of the European Union (case No. I ZR 140/15).

A tentative application of this framework (with some guidelines) to Service Providers in Research Infrastructures (like CLARIN) is beyond the scope of this abstract; for more details on this question (from the point of view of German, French and US law), see [Kamocki, 2014].

2.2. Liability of Service Providers in the New Directive on Copyright in the Digital Single Market

The new Directive on Copyright in the Digital Single Market (hereinafter: the DSM Directive) introduced a new category of Service Providers called ‘online content-sharing Service Providers’. They are defined as providers of services “of which the main or one of the main purposes is to store and give the public access to a large amount of copyright-protected works or other protected subject matter uploaded by its users, which it organises and promotes for profit-making purposes” (art. 2(6)). YouTube is a typical example of such a service.

The liability of online content-sharing service providers (OCSSPs) is subject to complex and much stricter rules (art. 17, several pages long). Because the DSM Directive is of recent vintage, there is no consensus yet on how to interpret these new rules, and their detailed analysis would greatly exceed the allowed length of this abstract; only some basic observations can be made here.

Under the new Directive, the acts of the OCSSPs qualify as communication to the public within the meaning of copyright rules (art. 3 of the Directive 2001/29/CE), and therefore the OCSSP is in principle required to obtain authorisation (a license) from the rightholder (or rightholders) (art. 17(1)). A license obtained by the OCSSP automatically (*ex lege*) covers subsequent communication to the public by the users of the service, provided that it is carried out for non-commercial purposes (art. 17(2)). From the user point of view, this seems to mean that anything found e.g. on YouTube can lawfully be shared for non-commercial purposes, which may potentially be interesting for creating language resources (although it remains to be seen how this will be implemented in national laws of the Member States). On the other hand, from the OCSSP perspective, especially those hosting content with multiple rightholders (e.g. language resources), the obligation to obtain a license will be difficult to fulfil. It is interesting to note here that the DSM Directive also allows Member States to introduce extended collective licensing mechanisms (art. 12), which could facilitate the process of obtaining licenses, but it is too early to say which Member States will adopt it, and how.

The liability limitation for hosting providers under the e-Commerce Directive does not apply to OCSSPs (art. 17(3)). If the OCSSP fails to obtain a license from rightholders, it is liable for copyright infringement, unless it demonstrates that it made ‘best efforts’ to obtain the license; to ensure that any content for which it obtained a specific notification from rightholders will not be available via its service; upon receiving a notification from rightholders, to act expeditiously to remove and/or disable access to the notified content; and to prevent future uploads of this content (art. 17(4)). This would probably require close cooperation with rightholders, sophisticated mechanisms of content notification with human review (art. 17(9)), as well as screening of uploaded content (which is why, during the adoption process, the opponents of this solution, originally in art. 13, referred to it as ‘censorship machines’ (Reda)). This may be seen as a contradiction of the rule of art. 15 of the e-Commerce Directive, which expressly states that service providers shall have no general obligation to monitor content; however, art. 17(8) of the DSM Directive expressly states the new framework “*shall not lead to any general monitoring obligation*”.

It is not yet clear what would constitute ‘best efforts’ that OCSSPs must make and demonstrate in order to avoid liability. It seems that the standard will be flexible, taking into account the size of the service, the target audience and the type of material uploaded by users, and the costs of implementation of preventive solutions (art. 17(5)). A CLARIN Centre, if it qualifies as an OCSSP (see below), would be held to a significantly lower standard than YouTube.

Importantly, art. 17 of the DSM Directive takes into account some copyright exceptions and limitations (art. 17(7)). For example, if a video was made and uploaded by a user within the limits of parody, then the content, even if notified by the rightholder, should not be removed. Limiting the user rights to rely on exceptions for parody and pastiche, as well as quotation, review and criticism would indeed seriously impair their freedom of speech. It remains to be seen how this will be implemented in practice: will the screening algorithms be able to tell that this ‘blacklisted’ content (i.e. content that they should prevent from uploading) in this particular case is used for parody or criticism purposes? If a human will review and assess the content, than according to which standards (given that parody cases are regularly decided in court)? How long will it take (in the world where, when it comes to criticism and parody, sometimes minutes matter)? Finally, the research exception is not listed among the exceptions that users of online content-sharing services can rely on, so it seems that research purposes (regardless of how broad they are in the applicable law) cannot be an excuse (at least not *a priori*) for uploading content in online content-sharing services.

Paradoxically this new framework, targeted at large international OCSSPs like YouTube, will likely provide them with a competitive advantage. The compliance with the new obligations would require very significant means — means that large OCSSPs have, but not necessarily their smaller competitors (despite the mechanisms introduced to protect startups — art. 17(6)). It remains to be seen how the new rules will transform the Web.

3. Possible Impact of the New Framework on CLARIN Centres

As demonstrated above, the new framework regarding liability of OCSSPs is particularly strict and complex, to the point of potentially having a chilling effect on online content-sharing services. At the same time, the previous framework related to the liability of hosting providers will continue to apply. It is therefore of great importance for the future of CLARIN and its centres to determine if they qualify as hosting providers, or as OCSSPs.

According to art. 2(6) of the DSM Directive, some categories of Service Providers are expressly excluded from the definition of OCSSPs. This is the case of non-for profit online encyclopaedias (such as Wikipedia), open source software developing and sharing platforms (such as GitHub), online marketplaces (such as OLX or even, arguably, Amazon) as well as “not-for-profit educational and scientific repositories”.

It seems that today most CLARIN repositories are indeed concerned by this last exclusion, and so they are not OCSSPs and still qualify for the liability limitation for hosting providers.

However, the situation becomes more complicated if the repository is used for some sort of commercial (for-profit) purposes, such as charging (even only some categories of users) for access. Sometimes it can indeed be very difficult to draw a line between what is ‘not-for-profit’ and ‘for-profit’, but crossing this invisible line may have huge consequences as far as liability is concerned. This needs to be taken into account before any potential commercial uses of CLARIN repositories are discussed, let alone put into practice.

The proposed paper invites a wide debate between CLARIN centres, the Board of Directors and legal experts on how CLARIN repositories should be organised in order to best fit within the existing legal framework regarding liability of service providers. Such a debate does indeed seem necessary.

Reference

- Reda, Julia, Official Blog, <https://juliareda.eu/eu-copyright-reform/censorship-machines/>, accessed 29.04.2019;
- Kamocki, Pawel, 2014. The Liability of Service Providers in e-Research Infrastructures. Killing the messenger? Proceedings of the 9th Language Resources Evaluation Conference, Reykjavik;
- Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market;
- Directive of the European Parliament and the Council on copyright in the Digital Single Market [unpublished as of 29.04.2019];
- European Parliament, Providers Liability: From the eCommerce Directive to the future, 2017. In-Depth Analysis for the IMCO Committee, available at: [http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA\(2017\)614179_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614179/IPOL_IDA(2017)614179_EN.pdf), accessed 29.04.2019.

The Extent of Legal Control over Language Data: the Case of Language Technologies

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

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

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

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

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

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

Irene Kull
University of Tartu
Estonia
irene.kull@ut.ee

Ga Gabriel Tavits
University of Tartu
Estonia
ga Gabriel.tavits@ut.ee

Age Värvi
University of Tartu
Estonia
age.varv@ut.ee

Abstract

The article aims to increase legal clarity concerning the impact of data containing copyrighted content and personal data on the development of language technologies. The question is whether legal rights covering data affect language models.

1 Introduction

The development of language technologies (LTs) relies on the use of language data (LD). Language data is often covered with several tiers of rights (copyright, related rights, personal data rights). The use of this kind of data can be based on consent or exemption model (for further discussion, see Kelli et al. 2015; Kelli et al. 2018).

The relevant issue here concerns the impact of data's legal regime on LTs. The question is whether legal restrictions applicable to data apply to the language technologies that are developed using them as well. The article aims to reduce the legal uncertainty regarding how far, in the pipeline of developing language technologies, the original copyright and personal data protection¹ regulations apply. If we take a recorded phone call, for instance, it is obvious that copyright and data protection apply to a copy of that recording. At the other extreme, it is equally obvious that they do not apply to the Voice UI (User Interface) of a new fridge, even though the latter was trained on a corpus containing the former. The line

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

¹ The GDPR defines personal data as “any information relating to an identified or identifiable natural person (‘data subject’)” (Art. 4 (1)).

where the original rights cease to apply has to be somewhere between these points, and it is vital for researchers and developers to know where.

2 From language data to language technologies

The development of data-driven/data-based language technologies contains:

1. Collection of raw data (written texts, speech recordings, photos, videos, etc.). These often contain copyrighted material and personal data. Their development usually does not involve any other activities than the actual recording, initial cleaning and sanity-checking of the data.

Dangers for both copyright and personal data can be very real: re-publication of copyrighted works, surveillance by governments or insurance companies, etc.

Almost impossible to anonymise or pseudonymise completely, so that it would become mathematically impossible to identify any persons or reproduce any significant portions of copyrighted works.

2. Compiling of datasets, or collections of data (raw text corpora like Google News, Common Crawl or OpenSubtitles, speech corpora like the Prague DaTabase of Spoken Czech, etc.). The above, but collected and organised with a specific criterion in mind (e.g. speech recordings on a specific topic by residents of a certain region in order to capture the accent of the region); these datasets usually come in such quantities that any individual piece of data constitutes a negligible part of the whole, and could in principle be removed without affecting the usability of the dataset.

For copyright and personal data purposes, not different from raw data². The main practical difference is that the sheer volume of data may make it technically difficult for an individual to become aware that their data has been included in the dataset.

Creation of a dataset often involves a nontrivial contribution in gathering, organising, indexing, presenting, hosting etc. of the data.

3. Creation of annotated datasets (POS-tagged corpus of written texts like the ENC17, syntactically parsed corpora like the Universal Dependencies treebanks, etc.). The above, augmented with some kind of analysis.

Again, not different from raw data in terms of copyright and personal data, although the copyright holders of the raw data and the annotations may be different. The annotation layers may be stored separately and may even have some use on their own, but normal practice is to process copies of the original data together with the annotation layers so that the resulting dataset contains all of the original data.

Creation of an annotated dataset includes analysis of the data, either manual, semi-automatic or automatic.

4. Models. Data products developed from some sort of processing on the above, but not necessarily containing the above, which try to *model*, i.e. represent or describe, language usage. Examples: dictionaries, wordlists, frequency distributions, n-gram lists like Google ngrams, pre-trained word embeddings like in Grave et al. 2018, pre-trained language models like in Devlin et al. 2018.

Creation of a model involves significant amounts of work, expertise and (computational) resources. Steps include at least creation and/or selection of the algorithm, implementation of the algorithm in software, hardware setup (may even include custom hardware development), hyperparameter optimisation, model validation.

In rare cases, some model types may be consumer products of their own (e.g. dictionaries). Mainly, however, models are used in downstream tasks to create other products.

² In fact, it can be argued that data-sets qualify for database protection (for further discussion, cf. Eckart de Castilho et al. 2018).

5. Semi-finished products (text-to-speech engine or a visual object detector) and **finished products** (talking fridge). Out of scope for the current analysis, because their status as original works should be beyond doubt.

3 The legal status of models

The focus of the article is on models. It is crucial to determine whether the use of data containing personal data and copyright content influences the subsequent utilisation of the model. Therefore, copyright and personal data regulations are analysed.

3.1 Copyright perspective

From the copyright perspective, there are three relevant issues. Firstly, whether copyright material is used. Secondly, if it is used then whether there is a legitimate ground for the use. Thirdly, how to define models themselves within the copyright framework.

To answer the question about the copyright law impact on models, the requirements for copyright subject matter should be briefly outlined. The main and long-established requirement is that of **originality**. Work is protected if, and only if, it is original. Therefore, the originality requirement defines the copyright status of the input data. Oddly enough, this general requirement was never defined in international treaties or European *acquis*³. The task to define the legal meaning of originality for copyright purposes was mainly taken by the Court of Justice of the European Union (CJEU). As was explained in the seminal decision in the *Infopaq* case (C-5/08), originality means the author's intellectual creation. Another important explanation in the *Infopaq* case was that an extract consisting of eleven words could constitute an original work. The Court has also explained that a single word cannot be regarded as original and protectable work.

In the context of the current research, the originality requirement is important from two different perspectives. First, if originality is missing, the pre-existing text contained in a dataset is not protected and can be used without authorisation. Therefore, even if parts of this text are reproduced in the model, they are not protected as well. Second, even if a text as a whole is original and, therefore, protected, the question remains, whether the fragments used in the model are original on their own. If they are not, then again, they can be used without authorisation. Thus, originality must be established not only concerning the original work but also as regards the parts used.

In addition, in its latest case law CJEU has underlined, that, besides originality, a work also must meet the second requirement in order to be copyright-protected, i.e. it "*must be expressed in a manner which makes it identifiable with sufficient precision and objectivity, even though that expression is not necessarily in the permanent form*" (C-310/17). Arguably, this requirement in practice will be present in the majority of cases, because the texts used for models normally are expressed in a fixed form.

We consider that the development of the model is done through a data mining activity, according to the definition of the Copyright Directive, which defines text and data mining as "any automated analytical technique aimed at analysing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations" (Art. 2 (2)).

To answer the question of whether models are copyright protected, by the previous section, we must establish whether they meet the requirement of originality also on their own (irrespective of the input dataset).

One of the criteria that can be used for assessing originality has to do with the degree of human intellect invested in the process: how far is the model a unique product, the result of the intellectual creation of the author (developer) and not the result of a process that any other qualified engineer could also create? Building a model (as presented in the previous section) includes a number of choices and actions on the part of the developer: choice/creation of the dataset, choice/creation of the programme to be used for the training and development of the model and various cycles of testing and validation by tuning the parameters of the training programme. So, the question is: if the same program is used by another qualified user (engineer) on the same dataset, would they arrive at the same results, i.e. produce the same model? The main differences lie in the tuning of the parameters of the algorithm/program

³ Although it was defined in several EU directives with regard to specific categories of works, such as computer programs or photographic works.

which is linked to the cycles of testing and validation; so, if this tuning is "original/creative" enough it can be considered a copyrighted program. If the choice of parameters is limited, as it happens in specific cases, the program and the model could not be considered original (cf. Eckart de Castilho et al. 2018). In other cases, this may indeed involve a substantial intellectual effort on the part of the developer; if so, it can be argued that the resulting program is "original" and thus copyright-protected. But what about the output of the processing, the model? Is this also "original"? The application of the same algorithm with the same parameters on the same input will result in the same model. In fact, this can be seen as similar to using a Part-of-Speech tagger to automatically annotate a corpus without any human intervention: the tagger itself may be copyrighted, the input dataset may also be copyrighted, but the annotations themselves cannot be considered as "original"; what is copyright protected is the part of the input dataset that remains in the annotated dataset. We can thus argue that the model itself is not original in this sense.

Also, we need to establish if any substantial original parts of the input dataset remain in the model and thus qualify for copyright protection of these parts. If none or very small parts remain in the model (and thus does not contain any original parts), then we can conclude that as far as this point is concerned, the model is not copyrightable.

In case considerable parts of copyrighted works remain in models, they can be considered derivative works. There is no clear definition of derivative work in international or European legal acts, and different jurisdictions have a quite different understanding of this concept (for further discussion, see Birštonas and Usonienė, 2013; Echart de Castilho et al. 2018). It is not clear, how much of the original work should remain to categorise a model as a derivative work. However, this issue is not very practical since the copyright protection of the primary work (copyrighted content in the dataset used for the development of the model) is not dependent on the fact whether the model is a derivative work or not. The only important question is whether the original part of the primary work has been used in the latter work (the model).

It is generally not possible to recreate copyrighted works or personal data contained in a dataset from the model that has used it. Some small excerpts of the original data may remain in the model, and it is important to see if these violate the regulations of copyright and personal data. Very idiosyncratic language use would be hard to filter out in a guaranteed way, but cases of this are very rare and lose significance even more with the increase of data volumes.

To give a definite answer, we should have a closer look into all the model types and the processes and resource types and modalities they have been built upon, which is not possible in the limits of this article. It can be argued though that models by definition try to capture *generalities* of language use and *abstract* from the original texts as far as possible, producing mainly lists of words or phrases and patterns with statistical measures.

3.2 Personal data perspective

Regarding personal data, it is theoretically possible that small but identifiable bits of information make it to the model. A wordlist might contain a name or e-mail address, for instance. This is easy to avoid using anonymisation or pseudonymisation.

However, it should be kept in mind that for personal data, there is no minimum segment in the audio synthesis. Even if the voice is synthesized using neural networks without any remnants of the person's original voice recording, having trained the network for research purposes using a publicly available radio transmission as training data, one is still using the personal data of that person when the person can be identified based on the synthesized output despite the fact that there is no single bit in the network which could be attributed to the person's voice.

The main issue here is how to substantiate the processing⁴ of personal data contained in a model. Generally speaking, the compilation of datasets containing personal data used to create models can be based on the consent, public interest research and legitimate interest (see, GDPR Art. 6 (1) a), e), f)). In

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

case there is consent to process data for research purposes, or processing relies on public interest and the resulting model is used for research purposes as well (it is not made available to the public or used for commercial purposes), then there is no problem. There is also no problem if consent covers commercial use and public dissemination.

However, the situation becomes complicated when a dataset containing personal data is processed based on consent asked for research or on the public interest research exception, but the resulting model (where the personal data may remain) is planned to be used for commercial purposes or be made publicly available. If the personal data is in the form of speech, then anonymisation is rather difficult. In the described case there are the following scenarios:

- 1) Argue that voice without any identifying information is not personal data (it is anonymous data). The key here is how to interpret the concept of identifiable natural person (Art. 4 (1))⁵;
- 2) Ask for consent for commercial use;
- 3) Argue that the use of voice in the model is based on the legitimate interest. Especially bearing in mind that the identification is impossible or almost impossible and the voice does not contain any data which would affect the data subject negatively.

4 Conclusion

It is clear that raw data, datasets and annotated datasets are affected by copyright and personal data regulations. To some extent, models rely on datasets. They do not usually contain copyright-protected content. However, models containing speech need to address personal data issues.

When creating a language model, there are several activities involving complex human intellectual activity such as choosing and annotating datasets as well as choosing the software and tweaking its parameters. The outcome of the preparatory software activities is applied to a prepared dataset to compile a language model.

The outcome of the preparatory software activities is usually encoded in a piece of software. This software becomes the model trainer that embodies the copyright of the preparatory software activities. The model trainer is applied to the prepared dataset but does not inject the copyright in the model trainer into the language model as the model trainer is a piece of software, which is mechanically applied to the dataset.

If the compiled language model contains sufficiently long pieces of the original data, there may be some copyright left from the dataset.

⁵ See e.g., Article 29 Working Party Opinion 4/2007 on the concept of personal data.

References

- Birštonas, R., Usonienė, J. 2013. Derivative Works: Some Comparative Remarks from the European Copyright Law. *UWM Law Review*, Volume 5.
- Case C-310/17. *Levola Hengelo BV vs Smilde Foods BV* (13 November 2018). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243886259&uri=CELEX:62017CJ0310> (14.4.2019).
- Case C-5/08. *Infopaq International A/S vs Danske Dagblades Forening* (16 July 2009). Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005> (14.4.2019).
- Directive (EU) 2019/... of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Text with EEA relevance). Available at <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2019-0231+0+DOC+XML+V0//EN&language=EN#top> (14.4.2019).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs].
- Eckart de Castilho, R., Dore, G., Margoni, T., Labropoulou, P. & Gurevych, I. 2018. A legal perspective on training models for Natural Language Processing. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, ELRA. Available at: <http://www.lrec-conf.org/proceedings/lrec2018/pdf/1006.pdf> (17.4.2019).
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. 2018. Learning word vectors for 157 languages. ArXiv Preprint ArXiv:1802.06893.
- GDPR. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679> (15.4.2019).
- InfoSoc Directive. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal L 167*, 22/06/2001 P. 0010 – 0019. Available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029> (14.4.2019).
- Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Chiara Kolletzek, Penny Labropoulou, Maria Gavriilidou. 2018. Processing personal data without the consent of the data subject for the development and use of language resources. *CLARIN Annual Conference 2018 Proceedings: CLARIN Annual Conference 2018, 8-10 October 2018 Pisa, Italy*. Ed. Inguna Skadin, Maria Eskevich. CLARIN, 43–48. Available at https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf (14.3.2019).
- Aleksei Kelli, Kadri Vider, Krister Lindén. 2015. The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: *Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wrocław, Poland*. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13–24. Available at <http://www.ep.liu.se/ecp/article.asp?issue=123&article=002> (28.3.2018).

User Support for the Digital Humanities

Tommi Antero Pirinen

Hanna Hedeland

Heidemarie Sambale

Hamburg Centre for Language Corpora (HZSK)

Universität Hamburg, Germany

{firstname.lastname}@uni-hamburg.de

Abstract

In this article we describe a user support solution for digital humanities. As a case study we show the development of the CLARIN-D helpdesk from 2013 into the current support solution that has been extended for DARIAH-ERIC as well as number of other non-CLARIN-D software and projects and describe a way forward for common support platform for CLARIN-DE that we are currently building towards as well.

1 Introduction

For both the ongoing digitalisation of humanities research in general and the CLARIN infrastructure in particular, the non-technical aspects of adequate training and user support are crucial for acceptance and involvement from the research communities. Many humanities researchers come from a rather non-technical background, and the use of digital tools and resources has not yet entered the curriculum to an appropriate extent. Researchers thus face various problems when confronted with tools and platforms for digital humanities research, many of which might not be predictable to the developers. Improving the usability of such kinds of tools and providing comprehensive documentation is of course very important, but in the end there is no replacement for a reliable help desk to assist users when they for various reasons struggle with digital resources, tools and services or need qualified advice in methodological questions.

Since digital humanities is multidisciplinary, the user support becomes a very central and important resource for exchanging information and experience, and for gathering expertise from various contexts. Apart from providing users with reliable support, the direct interaction with the users also provides valuable input about the users and their behaviour and on existing problems with the tools and platforms for infrastructure providers and developers. In our paper, we describe the development of a comprehensive resource based on a sustainable platform with re-usable workflows, for which we have also developed various strategies for scalability.

The rest of the article is organised as follows: in section 2 we describe the background of our approach to user support and our help desk, in section 3 we describe the current developments, in section 4 we introduce the scalability and in section 5 we describe the technical implementation of various challenges regarding management and scalability of the help desk, and finally in section 6 we describe our future plans based on the current development.

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

2 The CLARIN-D Helpdesk

The CLARIN-D Helpdesk (Lehmberg, 2015) was first launched in 2013 to provide the necessary user support for the emerging CLARIN-D infrastructure. After a thorough review of current ticketing systems, the open source OTRS platform¹ was chosen to meet the needs of the infrastructure's users and developers. In OTRS, the support requests are managed using support *tickets*. A ticket represents the entire conversation with the user and documents the interaction step by step. The tickets are organised into *queues*, which represent the various support areas, and the support is provided by *agent* users, who are active within one or more support areas, i.e. queues. For each ticket, the responsible agent and the owner of the ticket, who will be answering the inquiry, are defined and these roles become visible to other agents. Apart from the communication with the user, it is possible to communicate internally in a structured and documented manner within the ticket conversation to find an appropriate answer to an inquiry collaboratively.

This technical infrastructure and the related concepts and workflows comprise the relevant functionality required to reliably answer and document incoming queries. The use of a ticketing system solves most of the issues related to providing user support via email, e.g. relevant information (such as previous answers, templates etc.) for a group of people is not accessible due to the use of private mail accounts, or duplicate answers are sent when a common mail account is used by several persons and coordination fails. In any case, status and responsibilities for inquiries need to be managed without designated functionality.

As shown in figure 1, tickets arrive to the help desk from multiple sources. Depending on the origins, the tickets may need further sorting in the queues and assigning to the agents. The ticketing system records metadata about the issues, such as first response times and closing times, that are used to gather statistics relevant for the goal of providing efficient user support. The textual content of the help desk, i.e. questions and answers, can also be searched as a semi-structured knowledge base, or used as the basis for edited FAQ articles which are also distributed from the help desk.

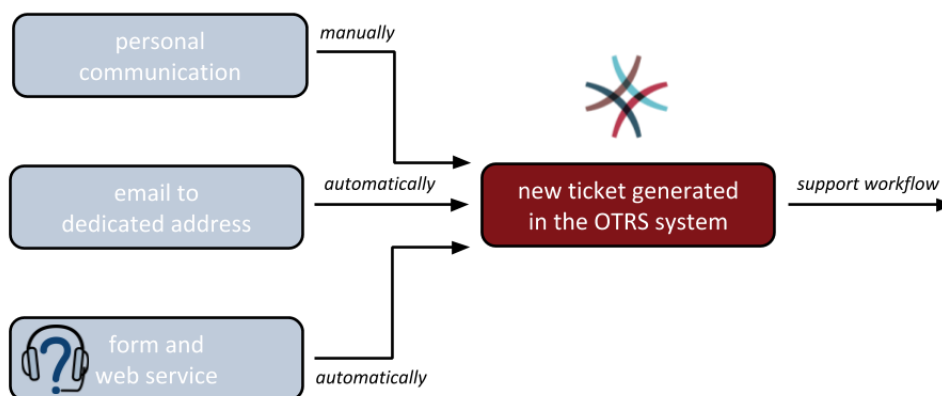


Figure 1: Tickets can be generated in various ways, allowing for presorting and delegation through queue-specific addresses or web service parameters.

Since the CLARIN-D infrastructure is distributed and centre-based, the queue structure of the CLARIN-D Helpdesk models the centres and the services they provide, and the help desk is used to distribute the support requests to the relevant experts. The tickets can be automatically sorted and delegated through the use of parameters from email addresses or web forms, or manually by the first line support agents as shown in figure 2. The first line support is carried out by experienced student assistants who receive specific training to be able to monitor and manage tickets, answer common and general questions, and delegate incoming inquiries that have not been sorted automatically. Administrators and

¹<https://otrs.com/>

first line support have a complete overview of all queues and agents whereas most expert agents of the second line support will only be able to see tickets and queues relevant to them. The experts comprising the second line support are researchers and developers of the participating centres and projects. Before an inquiry is successfully closed, several experts might contribute with their respective expertise on a complex matter, requiring the ticket to be moved across queues and reassigned several times. 95 % of the tickets are answered within two days.

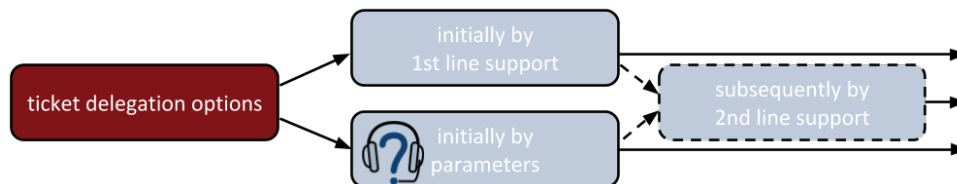


Figure 2: Tickets are sorted into queues and assigned to the right persons automatically or manually.

2.1 Distributed Support within CLARIN-D

Some support areas are not restricted to specific centres and some questions should better be answered from several perspectives. An example of an area often requiring coordination between multiple experts and centres is support for projects interested in creating digital spoken language resources using transcription software. Beyond phonetics, the creation and analysis of spoken language resources usually requires a great deal of technical and methodological support, both due to the mostly non-technical background of the users, and due to the highly complex nature of the task at hand. Spoken corpora comprise various interrelated data types and file formats and complex metadata valid across the corpus' components. For the creation of transcripts several software systems exist that are specialized for specific scenarios. Apart from expertise regarding certain tools, centres also provide expertise according to the raw data, e.g. depending on the language of the resource.

Based on the areas of expertise depicted in figure 3, a simple technical question regarding the transcription software EXMARaLDA might be initially answered by the first line support at the HZSK centre, the more complex question in the reply from the user looking into transcription software options would then forwarded to the second line support, who might find that this user should rather try the tools provided by the IDS or the BAS centre and forward the ticket accordingly.

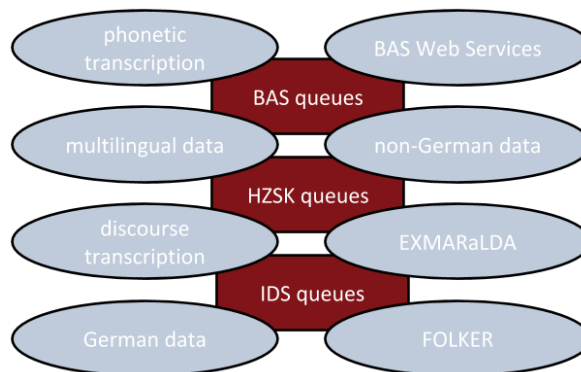


Figure 3: Professional support within the complex area of spoken language resources requires coordination and cooperation between several centres providing complementary expertise.

3 Support beyond CLARIN-D

Beyond CLARIN-D, the CLARIN-D Helpdesk is used to provide support on the European CLARIN level; for the CLARIN Virtual Language Observatory user feedback and outreach - including both generic feedback about user experience or reports of erroneous metadata - and for the CLARIN Federated Content Search Aggregator. For the VLO, the user feedback is divided into tickets regarding the VLO application and tickets regarding the metadata and the resources it describes. The tickets regarding the VLO application are handled by the VLO developers and the metadata related tickets can be handled by first line support or by the SCCTC Metadata Curation Taskforce as a part of their quality assurance work.

Apart from tools and services directly integrated into CLARIN, some support workflows for related tools from other contexts have also been successfully integrated into the CLARIN-D Helpdesk. This allows for interaction between developers and users of these tools across expertise and support areas and thus for a wider outreach. This could be a first step for these partners to become a part of the emerging Knowledge Sharing Infrastructure even though they are not yet certified CLARIN centres.

4 CLARIAH-DE - joining forces

As CLARIN-D and DARIAH-DE are merging into CLARIAH-DE, the support will also be merged into one comprehensive help desk based on the current CLARIN-D Helpdesk. One of the concepts we introduce in this article is scaling the help desk offerings to this larger user-base and fitting the workflows together, as will become necessary for user support in the CLARIAH-DE domain. Apart from branding and design related questions, this endeavour will require a thorough analysis of existing areas of expertise and a remodelling of queues and responsibilities. On the European level, we have already integrated the CLARIAH-ERIC help desk into our system, so that we are expecting to run a national and international CLARIAH help desk within the next couple of years.

5 Customizing ticket handling for increased flexibility

To be able to integrate support workflows for different tools and services from different contexts, we aim to make the help desk platform and all related components highly flexible. The support system is provided to the end users typically as a part of a web site or web application. For seamless integration of the support, there are few different technologies on which the web site managers can build a support integration. The most basic model of integration can be achieved with a support email address that forwards to the help-desk systems. The CLARIN-D Helpdesk also has a public web API based on SOAP that can be used to create and assign support requests. We provide a reference implementation of an HTML form for this API in our public GitHub repository that can be customised according to the requirements for the various support areas.² The integration of DARIAH-ERIC has been built on top of this API for WordPress and is likewise available at their GitHub repository.³ Each of the ticket creation workflows in the schematic presentation in figure 1 come with benefits and costs. Some of the technical challenges have been to enhance the automatic ticket assignment to minimise the overhead of the first line support in dealing with spam and easily categorisable tickets so their time can also be more efficiently used in actual user support tasks while keeping integration of further support areas as simple as possible.

6 Outlook

We have described a scalable help desk system with sustainable workflows introduced for CLARIN-D but already used far beyond its original designation. In the future we hope to extend the help desk and the areas of support catered for even further. Apart from the challenge ahead in merging the CLARIN-D and DARIAH-DE user support, we are looking forward to integrated complementary queues and support workflows and further enhance the usage of the help desk as a central knowledge resource for the digital humanities.

²<https://github.com/hzsk/clarind-helpdesk>

³<https://github.com/DARIAH-ERIC/contact-helpdesk>

References

Timm Lehmborg. 2015. Wissenstransfer und wissensressourcen: Support und helpdesk in den digital humanities.
In *FORGE*, pages 25 – 27.

CLARIN AAI and DARIAH AAI Interoperability

Peter Gietz
DAASI International
peter.gietz@daasi.de

Martin Haase
DAASI International
martin.haase@daasi.de

Abstract

Both CLARIN and DARIAH¹ have developed an Authentication and Authorization Infrastructure (AAI) which allows language and humanities researchers, respectively, to access on-line resources using their institutional accounts. Both AAIs are based on the SAML2 OASIS standard, and by virtue of this fact, lend themselves to interoperability. While CLARIN has established a "Service Provider federation" leveraging the German DFN-AAI federation and the international eduGAIN meta-federation, the DARIAH AAI has been built solely on top of eduGAIN via membership in the DFN-AAI, recently enhanced by the introduction of an IdP-SP Proxy for DARIAH services according to the AARC Blueprint Architecture. Both AAIs were successfully interconnected in 2018, already allowing many CLARIN and DARIAH users to access services from both communities today. However, there are some optimization possibilities that also are detailed in this paper.

1 Introduction

When it comes to cross-organizational authentication and authorization (AA), the SAML standard (Scott Cantor et al. 2005) is well-introduced. It allows for a clear separation between on-line services and institutions providing user data. The former role is called Service Provider (SP), while the latter role is the Identity Provider (IdP) in SAML terms. Almost all European countries, and many others worldwide, have established research and higher education trust federations in which services/SPs receive assertions from the research institutions/IdPs. The eduGAIN² meta-federation co-ordinates these national federations and provides for the exchange of such assertions beyond the boundaries of national federations. The actual trust framework builds on so-called SAML metadata, basically a collection of all server X.509 certificates of the instances of IdPs and SPs. From this point of view, a federation is nothing else but a list of server certificates managed by a federation operator with contractual processes for including new members.

For research communities consisting of multiple partners from different institutions and countries, this infrastructure is a perfect fit. The AARC project³ has analysed current research infrastructures and produced the so-called AARC Blueprint Architecture (AARC BPA. 2017) that helps research communities to interoperate with the national federations and eduGAIN. One key building block of the BPA is the IdP-SP-proxy: All services of a particular community simply need to connect to the IdP's component of the proxy, while the SP component of the proxy is the only communication partner for all IdPs in a federation.

Both, the CLARIN⁴ and the DARIAH⁵ research community, have built their respective AA infrastructures (AAI) upon the existing federations. This paper will introduce the two approaches and detail both the interoperability that has been achieved and potential future work in this area.

¹ We use the general names of these two humanities infrastructures, although most of the work reported here was done by the respective German national initiatives CLARIN-D and DARIAH-DE, since the work had been adopted by the respective ERICs on the European level.

² eduGAIN, a GÉANT initiative, see <https://technical.edugain.org/status>

³ Authentication and Authorisation for Research and Collaboration, see <https://aarc-project.eu>

⁴ Common Language Resources and Technology Infrastructure, see <https://www.clarin.eu>

⁵ Digital Research Infrastructure for the Arts and Humanities, see <https://www.dariah.eu>

2 Current DARIAH and CLARIN AAI Architectures

The **CLARIN** AAI (Dieter Van Uytvanck et al. 2017) initially had started as a separate federation, independent of eduGAIN. This CLARIN "Service Provider Federation" (SPF) contained at that time SAML metadata for

- the IdPs of all the CLARIN-affiliated institutions ("CLARIN centers")
- the SPs for all CLARIN services provided by these centers.

Thus this SPF existed alongside the established federations, did their own IdP discovery, own SAML metadata handling etc. The motivation was to convince the IdP operators to allow authentication and release the user attributes to all CLARIN SPs. CLARIN saw more benefits in directly reaching out to those IdPs that belonged to their research constituency and asking them to connect to the CLARIN SPF than trusting that eduGAIN would integrate them on the long term in such a way that the Sps could receive the needed attributes via that infrastructure.

When eduGAIN became well-established, i.e., when more institutions within the national federations opted-in to have their SAML metadata published in eduGAIN, CLARIN also allowed for directly publishing the SP metadata in the eduGAIN infrastructure. This means that SPs will be enrolled into eduGAIN by one of the participating national federations. This can be the national federation of the concrete CLARIN center offering some particular service, however, currently DFN-AAI is leveraged for registration of the vast majority of CLARIN SPs⁶. These SPs will be labeled as "clarin-member" by means of SAML metadata entity attributes. On the other hand, all CLARIN center institutional IdPs are integrated in eduGAIN as well, via their national federations. The SPF operators made sure that

- all IdPs in institutions with CLARIN centers, and
- further IdPs of universities with CLARIN customers

would release the proper set of user attributes to any SP that carries the *clarin-member* label. To achieve this, the SPF operators have approached the operators of these IdPs (henceforth "*CLARIN-enabled IdPs*") directly, instead of solely relying on initiatives like R&S or CoCo (see below).

In addition, CLARIN has set up a CLARIN Identity Provider (Dieter Van Uytvanck et al. 2017) in a later phase. It can be used by those potential users who have no access to an institutional IdP either because of the lack of a national federation, or because of no institutional affiliation (e.g. citizen scientists). This IdP is managed by the CLARIN ERIC⁷ with a best effort account verification.

The **DARIAH** AAI (DARIAH AAI) has been using eduGAIN from the start. It has also been operating an own IdP from the beginning for the same reasons as CLARIN, but also because even if the user had an institutional IdP, that IdP would not release a permanent and same identifier across several DARIAH SPs. Additional personal data, which even fewer IdPs released to DARIAH SPs for data privacy related reasons, needed to be requested from the user by the DARIAH infrastructure. The DARIAH IdP is following respective rules about on-boarding and off-boarding of users, that allows it to be a member of the DFN-AAI and to be connected to eduGAIN. These rules are a sophisticated vetting process that ensures all of its users are true members of the research community.

Almost all DARIAH services have also been part of eduGAIN right from the start. However, there is no such concept as a "DARIAH center", or "DARIAH IdP": users of any IdP that is part of eduGAIN can use DARIAH services. There is, however, a central registration step involved during authentication of such a user, such that, in addition to registering personal data like email address, agreement to DARIAH terms of use (ToU) and authorization group memberships can be handled. In its first form, the DARIAH AAI was a true "mesh" federation, i.e. each SP had to establish a 1:1 relationship with each eduGAIN IdP and had to implement an attribute query to the DARIAH IdP and possibly redirect to the central registration, such that information about ToU and groups were available.

When the AARC "Blueprint" became public, the DARIAH community was one of the first that introduced the IdP-SP proxy model, called "DARIAH AAI 2.0" (DARIAH AAI. 2018). Thus communication for both IdPs and SPs has become easier:

- eduGAIN IdPs only need to connect, and configure attribute release, for one single DARIAH SP, i.e. the proxy's SP component

⁶ As of 2019-08, eduGAIN has 42 CLARIN SPs registered via DFN-AAI, with a total of 46 CLARIN SPs.

⁷ ERIC: European Research Infrastructure Consortium: a European legal body to organize pan-European activities

- all DARIAH services only need to connect to the proxy's IdP component, and can be sure to receive ToU and group information, together with all required attributes without taking care of this themselves
- Furthermore, the proxy can handle central registration, releasing the individual DARIAH SPs of this duty.

As stated, **both** infrastructures faced the same problems in getting personalized data about the users from the IdPs and both now operate an own IdP and respective user management, so that any user can register an extra account there easily. While the CLARIN IdP is just a member of the SPF, the DARIAH IdP is a full member of eduGAIN, thus allowing authentication also for non-DARIAH services. Obviously the maintenance of an IdP and an user management as its base does imply regular effort, which needs to be sustained. Obviously it would make sense to combine efforts and only maintain one such infrastructure. Since in some countries, now including Germany, CLARIN and DARIAH have merged into one project ("CLARIAH"), this should be feasible.

The other view on this issue is attribute release from IdPs to SPs in general. CLARIN is solving the issue by maintaining a list of CLARIN-enabled IdPs. Only these IdPs are approached to release the required attributes, leaving aside well 95% of the remaining institutions that run an IdP in eduGAIN. For DARIAH, on the other hand, there is no restriction on IdPs. If an IdP would not release an identifying attribute about its users, this user was requested to contact their IdP administrator to make release possible for future users of this IdP. Complementing this, the DARIAH proxy of course has implemented two initiatives that allow for dynamic (i.e. not bilateral) attribute release from IdPs supporting them:

- the GEANT Data Protection Code of Conduct (CoCo. 2013)⁸
- the Research and Scholarship Entity Category (R&S. 2016)⁹

Adhering SPs would have the respective label (an entity category) in their SAML metadata. Supporting IdPs need only a single rule for attribute release, instead of having a separate rule for each SP, thus easing administration and enhancing user experience significantly. While R&S must be assigned by the federation operator to trusted SPs, and covers a fixed set of user attributes, CoCo can be self-asserted by an SP and must be accompanied by a PrivacyStatement document, and a list including the concrete user attributes this particular SP requests. CoCo has also been a requirement for CLARIN SPs lately.

Both infrastructures aim at all European research. While CLARIN was more successful in integrating services from many European countries, DARIAH also managed to integrate, besides all German DARIAH services, at least some important non-German services as well, e.g. the DARIAH-EU inkind contribution tool. For this cause, 5 DARIAH Service Provider Workshops had been organized, but only with small attendee numbers.

3 Interoperability Efforts

Since both infrastructures rely on the same technologies (SAML) and since both found ways to integrate with eduGAIN, it was rather easy to establish interoperability between them. Both infrastructures also co-operated from early phases onwards and also wrote a call for action on AAI advocating attribute release to these infrastructures (Call for Action. 2012) together.

As of the year 2018, the AAI interoperability between CLARIN and DARIAH has reached the status which is best illustrated in the following table. Note the remarks (*s), which are explained below.

Authentication for by	CLARIN Services	DARIAH Services* ^A
DARIAH "homeless" IdP	YES	YES
CLARIN IdP	YES	YES
eduGAIN IdPs* ^E	YES* ^C	YES* ^B
CLARIN-enabled IdPs	YES	YES* ^{B,D}

⁸ For newer activities see also <https://wiki.refeds.org/display/CODE/Data+Protection+Code+of+Conduct+Home>

⁹ For future activities see also <https://refeds.org/category/research-and-scholarship>

The following remarks must be made, since they will make an important distinction in the practical federation operation:

- *A: DARIAH Services: Almost all DARIAH-DE services and some other DARIAH services now authenticate via the proxy, meaning all eduGAIN IdPs can connect in principle.
- *B: Any eduGAIN IdP can access a DARIAH service if it has support for the entity categories CoCo or R&S, or if attribute release to the DARIAH proxy's SP half has been set up in a bilateral fashion, including at least eduPersonPrincipalName, eduPersonUniqueID, eduPersonTargetedID, or SAML2 persistent NameID.
- *C: an eduGAIN IdP can use a CLARIN service, if it has support for the entity categories clarin-member, CoCo, or R&S, or if attribute release to this CLARIN SP has been set up in a bilateral fashion, including the required attribute eduPersonPrincipalName, or eduPersonTargetedID
- *D: A DARIAH service can be used if this CLARIN-enabled IdP is registered in eduGAIN or has otherwise established bilateral trust with the DARIAH proxy's SP half.
- *E: There are no DARIAH centers that maintain their own IdP. However, DARIAH records a list of eduGAIN IdPs that have authenticated successfully once.

4 Future Work

Contributing to the past work, the following list of actions could be followed to make the CLARIN-DARIAH integration tighter. The actions are ordered from least to highest effort for highest to least impact. Some of the solutions (possibly (3), and (4) for sure) might be too complex to tackle since organizational and legal boundaries are involved.

- (1) Addressing *A: Within the DARIAH-EU Working Group FIM4D¹⁰, more and more DARIAH services will be connected to the DARIAH AAI.
- (2) Addressing *D:
 - It should be easy to set up the "clarin-member" entity attribute also for the DARIAH proxy. This ensures that at least CLARIN-enabled IdPs would send their attributes to all DARIAH services automatically.
 - For federations that have not implemented attribute release based on entity categories, CLARIN could approach these operators in order to register the DARIAH proxy, as it is done for regular CLARIN services.
- (3) Addressing *C and *E: DARIAH records a list of eduGAIN IdPs that have authenticated successfully at least once. The operators of these IdPs could be approached to set up "clarin-member" entity attribute support, or CoCo, or R&S, all of which would allow for dynamic attribute release to CLARIN services in a scalable way.
- (4) Addressing *B: IdPs in the CLARIN SPF that are not in eduGAIN should join eduGAIN via their national federations.

More long-term aims should be the inclusion of new SSO standards OpenID Connect and OAuth2. There are a number of technical solutions that provide for trans-protocol SSO, so that OIDC applications can be accessed via a SAML based authentication and vice versa. Such solutions, e.g. Gluu or Shibboleth, with its new OIDC plugin can be used also according to the proxy architecture as proposed by the AARC Blueprint Architecture.

While it does make sense to merge the CLARIN and DARIAH approaches, the usage of the emerging EOSC AAI should also be considered.

5 Conclusion

This article shortly described the authentication and authorization infrastructures of the two humanities related ERICs and reported on the respective interoperability. Using the same technologies and world-wide infrastructures, and cooperating on different levels, interoperability could be established, which allows most users to access the respective services. A number of work items have been pro-

¹⁰ <https://www.dariah.eu/activities/working-groups/fim4d-federated-identity-management-for-dariah/>

posed to allow for even more researchers to access services from both infrastructures. With the implementation of the AARC BPA by the DARIAH AAI, interoperability with other generic infrastructures, like with EGI¹¹ and in future with EOSC¹² can also be established.

References

- AARC BPA. 2017. AARC Blueprint Architecture Revision 1.0, Deliverable AARC-BPA-2017, 18-04-2017; see <https://aarc-project.eu/architecture/>.
- DARIAH AAI. 2018. Easy Federated Identity Management with DARIAH AAI 2.0, Poster presented at the DARIAH Grand Tour, Darmstadt, September 2018.
- Call for Action. 2012. CLARIN-D and DARIAH-DE 2012 Call for action on federated identity, https://www.clarin.eu/sites/default/files/clarin_dariah_call-for-action-aa1.pdf
- Jozef Mišutka. 2016. Robust SPF: workflow and monitoring, Deliverable CLARINPLUS-D2.2 , 2016-06-30
- CoCo. 2013. GÉANT Data Protection Code of Conduct, Version 1.0, 14 June 2013, <http://www.geant.net/uri/dataprotection-code-of-conduct/v1>
- DARIAH AAI (without year) On-line documentation, <https://wiki.de.dariah.eu/display/publicde/DARIAH+AAI+Documentation>
- R&S. 2016. REFEDS Entity Category: Research and Scholarship v1.3, 8th September 2016, <https://refeds.org/wp-content/uploads/2016/09/ENTCAT-RANDS-v1.3.pdf>
- Scott Cantor et al. 2005. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0, OASIS Standard, 15 March 2005, <http://docs.oasis-open.org/security/saml/v2.0/saml-core-2.0-os.pdf>
- SPF (without year) On-line documentation, <https://www.clarin.eu/content/service-provider-federation>
- Dieter Van Uytvanck et al. 2017. Service Provider Federation Full Extension, Deliverable CLARINPLUS-D2.7, 2017-04-31

¹¹ European Grid Initiative, see <https://www.egi.eu/>

¹² European Open Science Cloud, see <https://www.eosc-hub.eu/>

Word at a Glance – a Customizable Word Profile Aggregator

Tomáš Machálek

Faculty of Arts

Charles University, Prague, Czech Republic

tomas.machalek@ff.cuni.cz

Abstract

Word at a Glance is a highly customizable web application serving as a word profile generator aggregating a diverse set of possible data resources and analytic tools. It focuses on providing means for expert-based interpretation and presentation of the data and, at the same time, it makes the results easily accessible to general public.

1 Introduction

CLARIN centres and other language infrastructures all over the world have already collected a vast amount of diverse, well organized language resources (LRs) for many languages. However, such a rich set of resources may be overwhelming and hard to navigate through for some users. Our experience in the Czech National Corpus (CNC) also shows that it is helpful to point users towards a careful selection of (sub)corpora, as well as to offer use cases with appropriate reasoning and interpretation of the observed language phenomena.

The aim of this paper is to introduce Word at a Glance (WaG) that has been created to tackle the problem. WaG is a web application that offers a brief profile of a query (typically a single word) that is attractive, user-friendly and easy to understand. It is designed as a set of customizable tiles to showcase relevant characteristics of a given word (or phrase) that can be derived from available data and presented to the user (visualizations of metadata variations, development trends, collocations, translation equivalents, variation of individual forms in a paradigm based on real data etc.).

We are aware of several alternative approaches that served as an inspiration for WaG. Most prominently, there is the DWDS online dictionary (Klein and Geyken, 2010) that is very similar, but lexically oriented. Another source of inspiration was Mark Davies's site English-Corpora.org (Davies, 2008) with its corpus-based overall statistics, as well as Sketch Engine (Kilgarriff, 2014) with its collocation profiles. However, these alternative approaches are not directly usable either because the interfaces are tailor-made for particular data or simply because of licensing reasons. The layout of WaG can be seen on Figure 1, the latest development version of WaG is running and can be tested at <https://trost.korpus.cz/wdglance/>.

2 Development considerations

From the application design point of view, the motivation introduced in the previous chapter leads to a web application which queries both local and remote language-related data resources of various types and transforms their responses into a set of predefined text and graphic representation types. Among considered data resources there are corpus search engines, relational/XML/NoSQL databases, and general web services providing machine-readable responses.

On the implementation level, such data aggregation is a common software problem and can be accomplished in many different ways considering both server and web browser environments. The most prominent property in this case is the concurrent nature of the application in terms of program execution. WaG queries many resources almost simultaneously and starts to render results as soon as they arrive (unless a specific dependency between data views is defined).

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

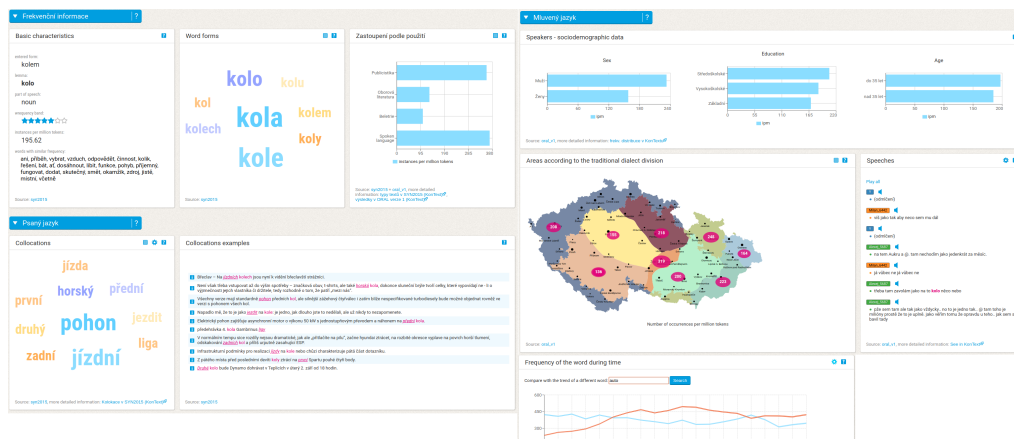


Figure 1: Screenshot of the application (slightly rearranged due to space limitations)

In the search of balance between project's generality and feasibility, we took a practical, bottom-up approach for the development of WaG by starting with services developed or used by the CNC (Machálek and Křen 2013, Škrabal and Vavřín 2017, Cvrček and Vondříčka 2011). Such a solution brings an unavoidable initial bias towards a certain set of services but on the other hand, WaG development is fast due to our knowledge of those services and we can also benefit from feedback given by our advanced users who are familiar with existing CNC applications and who can instantly compare their current experience against an existing WaG prototype. Also, KonText – our key corpus search application is developed with concurrent access in mind as it is able to distribute load across many different workers. Adopting of other resources and services is either in development (Clarín FCS, MySQL, SQLite) or considered (Corpus Workbench, ElasticSearch).

WaG is designed as a building kit where an administrator decides which data views will be installed along with which data resources are attached to those views. Unless a completely new type of visualization or data resource is needed, the administrator will be able to create a custom word profile page without any programming experience just by editing a documented JSON configuration file (Figure 2).



Figure 2: A sample configuration of two tiles along with their placement into the layout

It is important to point out that with such a configuration flexibility comes also a bigger responsibility of the administrator when considering individual visualizations, used data and added descriptions so that the users would not draw incorrect conclusions based on insufficient or skewed data.

3 Core concepts & terms

3.1 Operation modes

As a result of collection and analysis of use cases from existing CNC applications, we have defined the following operation modes:

- single word search (within a single language),
- two words comparison (within a single language),
- search for a word translation or a matching translated text chunk (two languages).

For the time being, this use case distinction represents an integral application design property of WaG and thus cannot be avoided when planning a concrete WaG deployment or when creating new data visualizations. However, it is not necessary to offer all the modes.

3.2 Tile

Tile is the basic building block of both WaG user interface and its inner architecture. It encapsulates a visual presentation of a specific language phenomenon along with necessary logic for fetching and transforming data from a configured resource. The WaG project already provides a set of operational tile types:

- concordance (including parallel variant for two aligned languages),
- concordance merged from multiple concordance subsets derived via additional filtering,
- word frequency profile,
- collocation profile,
- pie chart & bar chart for text metadata frequency information
- multi-source frequency bar chart (e. g. data merged from two or more different corpora),
- geographic area information mapped to an SVG image,
- time distribution of a text metadata attribute (e. g. publication date),
- general word translation and side-by-side comparable word translations based on multiple data subsets used to infer the translation.

These tile types are part of the core WaG project and their number is likely to grow by the time of the official public release. We are also considering ways how to involve third-party tiles.

Optionally, a tile can offer an **alternative view mode** which is typically a table view for a chart-based tile. Also, so called **tweak mode** is available for some tiles where user can change some easy to understand parameters – e. g.:

- in the collocation profile tile, both/left/right search context can be selected,
- time distribution tile allows attaching a trend of another word for direct comparison.

3.3 Connected tiles & sub-queries

While most of the tiles work independently on each other, the development and testing revealed interesting possibilities on how to connect relevant tiles to obtain more illustrative result presentation. Some tiles provide results which can be easily transformed into queries themselves. E. g. a tile providing collocation profiles and the translation tile give basically a list of words (along with some statistical data). This led to a concept of **sub-query**. We can take those words and run a “second query round” to which some of the tiles may respond. For example, each line in the concordance tile can be used as an example of the actual use of the given collocation from the collocation tile.

3.4 Tile data service

Each tile must have a configured data resource client (or multiple clients in some cases) which provides a way to obtain data from the resource. It is expected that a typical tile will be able to utilize different clients for the same function. E. g. the concordance tile is able to read data from KonText / NoSketch Engine (Rychlý, 2007) or from a Clarin FCS endpoint, the bar chart tile is able to read data from KonText, from an Elasticsearch instance or from an SQL database.

4 Implementation and development notes

WaG is implemented as a pure JavaScript application (for both client and server) with most of the functionality running in a web browser and some support functions running on a server. In our case, a typed variant of JavaScript called **TypeScript** has been used.

The WaG project is open-source with development coordinated via a GitHub repository at <https://github.com/czcorpus/wdglance>. In this phase, the project still evolves rapidly so it is possible that some programming interfaces and configuration directives will change. A running instance of the latest development version can be found at <https://trost.korpus.cz/wdglance/>. The public release is planned for the beginning of the summer 2019 which means a full presentation can be expected during the CLARIN Annual conference 2019. For both potential contributors and administrators, it is already possible to download and install a working prototype. We welcome any comments, bug reports and pull requests.

5 Conclusion

WaG is intended for use in language infrastructures that offer many language resources of various kinds and from different domains. Therefore, it has been designed with reusability in mind. Although its primary purpose is now to serve as a part of the CNC web portal, its architecture makes it possible to deploy WaG also by other projects. It can be easily embedded into existing web pages or run as a standalone website.

WaG is intended both for newcomers and advanced users. The newcomers can easily see the power of the data, while the advanced users appreciate it as an overview of typical word's behavior. They can then continue exploring the feature in question by simply clicking the tile that brings them to the respective application for further details, often with the possibility of modification of the original query. Therefore, WaG can also be seen as a starting point to the exploration of the variety of available corpora. The general message is: wherever you go, just start with WaG that may show you something interesting you didn't know about. We believe this could greatly improve not only the usability of the infrastructure, but also to attract potential users and to bring them closer to the language tools and technologies.

Reference

- Wolfgang Klein, Alexander Geyken. 2010. *Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften, <<https://www.dwds.de/>>, abgerufen am 01.04.2019. <https://www.dwds.de/>
- Adam Kilgarriff et al. 2014. *The Sketch Engine: ten years on. Lexicography*, 1: 7-36. <http://www.sketchengine.eu/>
- Petr Rychlý. 2007. *Manatee/Bonito - A Modular Corpus Manager*. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno : Masaryk University, p. 65-70. ISBN 978-80-210-4471-5.
- Michal Škrabal, Martin Vavřín. 2017. *The Translation Equivalents Database (Treq) as a Lexicographer's Aid*. In: I. Kosem et al. (eds): *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Lexical Computing CZ, s. r. o., Leiden, s. 124–137.
- Tomáš Machálek, Michal Křen. 2013. *Query interface for diverse corpus types*. In: *Natural Language Processing, Corpus Linguistics, E-learning*, pp. 166–173. Lüdenscheid: RAM Verlag. ISBN 978-3-942303-18-7.
- Mark Davies. 2008. *English-Corpora.org*, available from WWW: <http://corpus.byu.edu/>
- Václav Cvrček, Pavel Vondříčka. 2011. *Výzkum variability v korpusech češtiny*. In: František Čermák (ed.): *Korpusová lingvistika Praha 2011. 2. Výzkum a výstavba korpusů*. NLN, Praha, s. 184–195.

Technical Solutions for Reproducible Research

Alexander König
Eurac Research, Italy
Alexander.Koenig@eurac.edu

Egon W. Stemle
Eurac Research, Italy
Egon.Stemle@eurac.edu

Abstract

In recent years, the reproducibility of scientific research has more and more come into focus, both from external stakeholders (e.g. funders) and from within research communities themselves. Corpus linguistics and its methods, which are an integral component of many other disciplines working with language data, play a special role here – language corpora are often living objects: they are constantly being improved and revised, and at the same time, the tools for the automatic processing of human language are also regularly updated, both of which can lead to different results for the same processing steps. This article argues that modern software technologies such as version control and containerization can address both issues, namely make reproducible the process of software packaging, installation, and execution and, more importantly, the tracking of corpora throughout their life cycle, thereby making the changes to the raw data reproducible for many subsequent analyses.

1 Introduction

While reproducibility has always been one of the main pillars of scientific research, within the last ten years this has come even more into focus for the social sciences and humanities (SSH). Prominent cases of scientific fraud, for example the case of Diederik Stapel in the Netherlands (Levelt et al., 2012), have brought problems about the reproducibility of scientific research into focus. In this article, we discuss some possible techniques to handle this problem using standard tools from the realm of software development. We propose to use versioning software to ensure the persistence of data (see section 2) and containerisation to ensure the same for NLP tool-chains (see section 3). We will then briefly discuss a case study where this approach has been partly implemented (see section 4) and highlight some challenges that were encountered along the way (see section 5).

2 Ensuring persistence of data

After initially collecting the data of a (text) corpus it is common that the corpus keeps evolving while at the same time the first analyses are already being carried out. It is also likely that while working on the corpus and analysing the data, mistakes in the transcription or the annotation are discovered, which need to be corrected. And with a rich annotation scheme that is constantly being re-evaluated and refined this is usually all the more true.

While these kind of changes are unproblematic as long as the corpus is still in its "building phase", as soon as the first analyses have been made public, any change to the data will endanger the possibility of reproducing these analyses. Therefore, the researchers have to preserve a version of the corpus as it was when a specific analysis was made.

If the corpus in question is a text corpus (probably being stored in some kind of XML format like e.g. TEI¹), an obvious solution to this problem is to use existing versioning tools like subversion² or

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

¹<https://tei-c.org/>

²<https://subversion.apache.org/>

git³ to keep track of all changes within the corpus. This is also possible for corpora that are not mainly text-based, for example, multimodal corpora. First, they often have a text-component in their annotations (which could have been done, for example in ELAN⁴ or EXMARaLda⁵, both of which store their data in XML format) and second, the problem of storing large files in versioning tools is being addressed - and will likely eventually be solved. Using such an existing versioning software solution, all changes throughout the life cycle of a corpus can be tracked and through the use of code hosting platforms like Github⁶ or GitLab⁷ all changes can be made transparent to the research community as a whole.

But having the corpus available on such a code hosting platform, while having the advantage of being very transparent about all changes made to the data, might not be the ideal way of providing the data to other researchers. Therefore traditional data repositories like CLARIN Centres, META-SHARE⁸, and zenodo⁹ will still play a role in making the data available to the users and especially in providing findability (through participation in search interfaces like the VLO¹⁰ or OLAC¹¹) and issuing persistent identifiers to specific versions.

3 Methods and tools and their impact on reproducibility

In linguistic research – especially in the sub-fields of corpus linguistics and natural language processing – data is often processed with the use of quite intricate software tool-chains. Ranging from more simple tasks like tokenisation or lemmatisation to more complicated ones like fine-grained syntactic parsing. The unification of all the necessary tools for the automatic processing of human language into a unified processing framework is more the exception than the norm. This inevitably leads to a wide variety of individual solutions each with their own installation procedures, development life cycles with maintenance and update schedules, etc. (Wieling et al., 2018). Furthermore, linguistic models that are often at the heart of such tools are also subject to change, and this change need not necessarily be synchronized with the tool itself, spanning an even wider range of possible combinations (see e.g. (Nothman et al., 2018)).

This short overview already shows the difficulty for other researchers to exactly recreate a certain tool-chain to verify research results. It can only be ensured if the original researchers document their setup carefully, noting down exactly which version of a certain tool was used and how exactly the various tools were combined. An additional problem is that some software manufacturers do not make older versions of their products easily available, so even if the version is known it is not certain that it can be obtained when necessary. For this reason it has been discussed whether scientific software should be archived in research repositories alongside the data, but so far, while some CLARIN repositories do also host linguistic tools, little progress has been made in this regard.

The recent trend in software deployment and administration towards containerisation of services seems to us to be a promising solution to the aforementioned problems regarding the reproducibility of data processing in linguistic research.

Containerisation means that certain programs are installed not on a real computer or even a full-blown virtualised environment like a virtual machine, but instead in a very basic environment that leaves out everything that is not vital for the program in question to work. The idea is to minimize both the amount of memory and processor time needed for such a containerised service and also the possibility for unwanted side effects. With Docker¹² this way of packaging programs and services has been widely adopted within the last years and the additional possibility to orchestrate the deployment of such minimal

³<https://git-scm.com/>

⁴<https://tla.mpi.nl/tools/tla-tools/elan/>

⁵<https://exmaralda.org/en/>

⁶<https://github.com/>

⁷<https://www.gitlab.com>

⁸<http://www.meta-share.org>

⁹<https://zenodo.org>

¹⁰<https://vlo.clarin.eu/>

¹¹<http://search.language-archives.org>

¹²<https://www.docker.com/>

containers using a platform like Kubernetes¹³ makes using existing containers "off the shelf" quite easy, especially because a lot of the big infrastructure providers (e.g. Google¹⁴, Microsoft¹⁵ or Amazon¹⁶) offer ways to deploy containers on their infrastructure for a moderate price and there seems to be a trend to make this kind of deployment as easy as possible¹⁷.

As a researcher, building the tool-chain for a new project can be done directly in Docker. There are already a variety of places where the resulting docker images (from which various container instances can be created) can be stored for re-use by others. For example, GitLab offers such a Docker image registry for free. GitLab is also a place where the data can be stored (see section 2), and both the data and the tool-chain used could thusly be stored in one place, making it much easier for researchers planning to recreate an experiment to get both in exactly the same versions that had been used originally. The wide availability of container hosting (see above) also means that it will be quite easy to simply take such a container with the whole tool-chain setup and use it to verify the results or look for something else in another set of data while ensuring that the same methodology is used as in the original research.

4 Case Study: The MERLIN Corpus

The Institute for Applied Linguistics (IAL) at Eurac Research is currently investigating how it can move towards such a setup for more reproducibility in research as outlined in the previous sections. One of the first corpora that was transformed into such a strictly versioned environment is the MERLIN corpus (Boyd et al., 2014). The corpus is completely available on a publicly reachable on-premise GitLab installation¹⁸. The repository is divided into multiple parts for the various formats in which the data is available and is accompanied by extensive documentation. The different versions of the corpus are realised as tags in GitLab, while these tagged versions are also uploaded into the Eurac Research CLARIN Centre (ERCC), the CLARIN DSpace repository hosted by the IAL, so they can be easily downloaded by less tech-savvy users¹⁹. Another advantage is, of course, that this integration of the data into a CLARIN Centre will make the metadata available to various search engines (e.g. the VLO or the OLAC search) and it can therefore be discovered easily. All the data for a tagged version is available both at the ERCC and on GitLab with each of these hosting platforms referencing the other. At both places, all versions are accompanied by a changelog that explains the changes between versions. On GitLab, the interested user can also make use of the integrated version diff to get more fine-grained information on the changes between versions.

5 Challenges and Pitfalls

While trying to implement the paradigm as described above, we already encountered a number of surprisingly challenging cases than the ideal one of a corpus that can be provided completely as open access. As linguistic corpora always consist of personal data produced by individuals there are both privacy and IPR concerns that need to be considered. And if not all of the data can be made publicly available, there has to be additional access protection both on the side of the DSpace repository and on the side of GitLab. While it is easy to have some data require a login with an academic account (using the CLARIN federated login) in DSpace, the GitLab repository should ideally not be made completely password protected, but have at least an openly available landing page that describes the corpus. We have tried to implement this for the DiDi corpus (Frey et al., 2015) using git submodules where the main repository with the documentation and the overview of the various data formats is publicly accessible and the actual data is in sub repositories that require a login²⁰. It is likely that more complex access scenarios will prove even more difficult to map to a code hosting platform.

¹³<https://kubernetes.io/>

¹⁴<https://cloud.google.com/kubernetes-engine/>

¹⁵<https://azure.microsoft.com/en-us/services/kubernetes-service/>

¹⁶<https://aws.amazon.com/containers/>

¹⁷<https://cloud.google.com/blog/products/serverless/introducing-cloud-run-button-click-to-deploy>

¹⁸<https://gitlab.inf.unibz.it/commul/merlin-platform>

¹⁹<https://hdl.handle.net/20.500.12124/6>

²⁰<https://gitlab.inf.unibz.it/commul/didi/data-bundle>

There is also, as always when using external services for sensitive data, the consideration whether one should store their data with a commercial provider, possibly one based in another country and jurisdiction. One way to avoid this is the GitLab “community edition”²¹ that can be installed on local infrastructure, meaning that the researcher/the institute will be able to keep full control of the data and container hosting.

Another possible pitfall is foreseen in the use of Dockerfiles²² to create a persistent setup of a tool-chain. Unfortunately when writing the Dockerfile there is currently no enforcement of explicit versioning of the used software. This means that two containers built using the same Dockerfile at two different points in time can contain two slightly different versions of the software which may result in different behaviour. This has to be kept in mind when creating the Dockerfile and can be averted by always requiring explicit versions of a tool, for example by using the functionality that is provided for this by the various Linux package managers.

6 Conclusion and Outlook

Reproducibility of corpus-linguistic research is a central problem within the linguistic community (Wieling et al., 2018) which is currently not well addressed in a large number of projects. In this paper, we have presented a promising approach to tackle this problem using existing tools from the software development world. We have started using this approach for existing corpora, but already encountered a number of potential problems of which we highlighted some. Nevertheless, this way of ensuring that results in corpus-based linguistic research can easily be reproduced by fellow researchers seems like an idea that is worth pursuing in the future. It makes sense for the CLARIN community with its focus on providing infrastructure for improving the process and the outcomes of research to follow up on this and see how it can help researchers to make their research easier to reproduce for others. Maybe CLARIN can offer a central infrastructure for the hosting of code/data on the one hand and docker images on the other, for example by installing a gitlab instance on one of the CLARIN-ERIC servers. Another possibility would be to develop guidelines or best practices for this kind of setup that can then be followed by the CLARIN community and which would result in a distributed infrastructure for version managed data and processing tool-chains.

References

- Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Stindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *LREC*, pages 1281–1288.
- Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W Stemle. 2015. The DiDi corpus of South Tyrolean CMC data. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media at GSCL2015 (NLP4CMC2015), Essen, Germany*, pages 1–6.
- Willem JM Levelt, PJD Drenth, and E Noort. 2012. Flawed science: The fraudulent research practices of social psychologist diederik stapel.
- Joel Nothman, Hanmin Qin, and Roman Yurchak. 2018. Stop Word Lists in Free Open-source Software Packages. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 7–12, Melbourne, AU, July. Association for Computational Linguistics.
- Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Reproducibility in Computational Linguistics: Are We Willing to Share? *Computational Linguistics*, 44(4):641–649, December.

²¹<https://about.gitlab.com/install/ce-or-ee/>

²²the recipe for constructing a container, see <https://docs.docker.com/engine/reference/builder/>

Approaches to Sustainable Process Metadata

Kerstin Jung and Markus Gärtner

Institut für Maschinelle Sprachverarbeitung

University of Stuttgart, Germany

{kerstin.jung,markus.gaertner}@ims.uni-stuttgart.de

Abstract

Metadata of a resource is information about a resource but not part of the resource itself. However, providing metadata is a crucial aspect of resource sustainability. In this contribution we show examples how to collect and provide additional process metadata, i.e. data on the creation process of a resource and decisions made in this process, to further increase the value of a resource.

1 Introduction

Nowadays sustainability of resources and results is a key feature with respect to research based on (language) data. Major aspects of sustainability are (i) reusability, (ii) reproducibility and (iii) comparability, where the first means not to redundantly repeat work already done by others, e.g. in not creating resources which already exist, the second means to be able to repeat work done by others, e.g. to show that results are valid, and the third means to be able to relate own findings to the findings of others. While (i) emphasizes a cost factor, (ii) and (iii) are traditional devices in academic discourse.

In addition, in a time of big data collections on the one hand, and a rich set of very specific small speech and text corpora on the other hand, reusability takes a twist back to more interdisciplinary uses. A corpus of computer mediated communication can be of use in the social sciences as well as for linguistic research. A collection of concurrent editions of a specific drama can be of use in literature studies as well as in regional history. This is where metadata and process metadata come into play.

Metadata, such as information about language, authors, time coverage or data format constitutes a set of features by which an existing resource can be found by potential users from various disciplines. In addition, process metadata, such as information about tools used and decisions made in the creation process allows users to decide if a resource fits their needs. For example, a researcher interested in spelling might want to know if and how a normalization of spelling “mistakes” has taken place in a corpus.

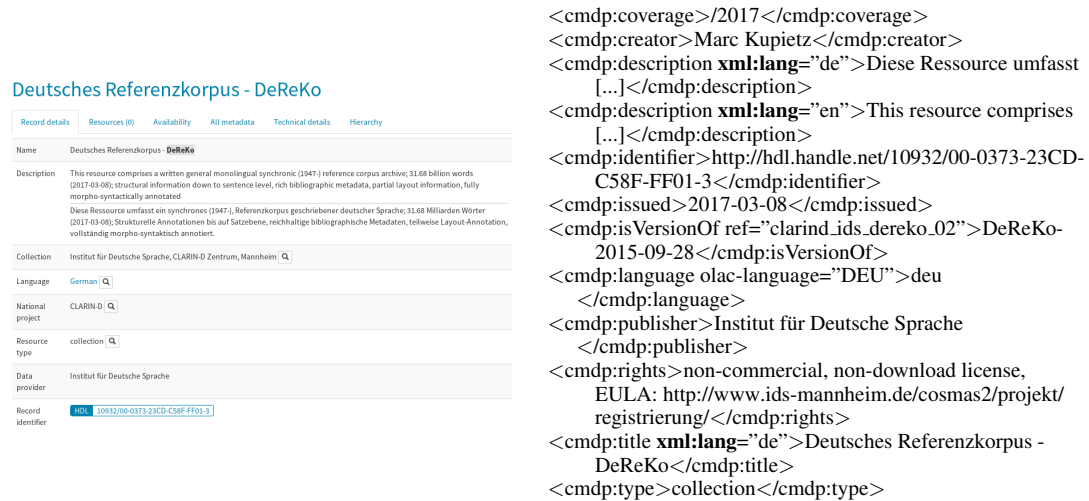
One aspect of reusability is also availability but not all resources can be made available to the community, e.g. due to protection of the participants in recordings or due to specific copyrights. However, since metadata is data about the resource and does not contain data from the resource itself, at least the metadata can be published, such that other researchers are able to estimate the comparability of their findings on a data set, e.g. with respect to size or temporal coverage, or the validity of a dataset for a specific study. Knowing that a resource exists can also be a starting point for new collaborations, in which data can often be shared even if it cannot be made available to a greater range of researchers.

In the following we describe two established CLARIN examples for metadata and process metadata and present the handling of process metadata within a corpus project and a tool to support the creation of process metadata in workflows of the Digital Humanities. Thereby we apply a broad definition of the term *resource*, comprising corpora and collections but also tools, lexical knowledge bases and other items appearing in the context of language data research.

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

2 Static Metadata vs. Process Metadata

We distinguish classical metadata from process metadata. Regarding language resources typical metadata schemas are Dublin Core¹ with 15 core elements including *title*, *coverage*, *creator*, *language* and *format* and the TEI Header². Catalogs like LDC³ or the LRE Map⁴ apply small sets of dedicated metadata features to describe and search for the listed resources. In the context of CLARIN, the flexible Component MetaData Infrastructure (CMDI)⁵ has been established along with the Virtual Language Observatory (VLO) catalogue⁶ (van Uytvanck et al., 2012; Goosen and Eckart, 2014). Figure 1 shows an entry in the VLO together with an excerpt of the underlying metadata.



The figure shows a screenshot of the VLO entry for 'Deutsches Referenzkorpus - DeReKo' on the left and an excerpt of its underlying CMDI metadata on the right.

Deutsches Referenzkorpus - DeReKo

Record details | Resources (0) | Availability | All metadata | Technical details | Hierarchy

Name: Deutsches Referenzkorpus - **DeReKo**

Description: This resource comprises a written general monolingual synchronic (1947-) reference corpus archive; 31.68 billion words (2017-03-08); structural information down to sentence level, rich bibliographic metadata, partial layout information, fully morpho-syntactically annotated. Diese Ressource umfasst ein synchrones (1947-) Referenzkorpus geschriebener deutscher Sprache; 31.68 Milliarden Wörter (2017-03-08); Strukturelle Annotationen bis auf Satzebene, reichhaltige bibliographische Metadaten, teilweise Layout-Annotation, vollständig morpho-syntaktisch annotiert.

Collection: Institut für Deutsche Sprache, CLARIN-D Zentrum, Mannheim

Language: German

National project: CLARIN-D

Resource type: collection

Data provider: Institut für Deutsche Sprache

Record identifier: [10932/00-0373-23CD-C58F-FF01-3](https://hdl.handle.net/10932/00-0373-23CD-C58F-FF01-3)

```
<cmdp:coverage>/2017</cmdp:coverage>
<cmdp:creator>Marc Kupietz</cmdp:creator>
<cmdp:description xml:lang="de">Diese Ressource umfasst
[...]</cmdp:description>
<cmdp:description xml:lang="en">This resource comprises
[...]</cmdp:description>
<cmdp:identifier>http://hdl.handle.net/10932/00-0373-23CD-
C58F-FF01-3</cmdp:identifier>
<cmdp:issued>2017-03-08</cmdp:issued>
<cmdp:isVersionOf ref="clarind_ids_dereko_02">DeReKo-
2015-09-28</cmdp:isVersionOf>
<cmdp:language olac-language="DEU">deu
</cmdp:language>
<cmdp:publisher>Institut für Deutsche Sprache
</cmdp:publisher>
<cmdp:rights>non-commercial, non-download license,
EULA: http://www.ids-mannheim.de/cosmas2/projekt/
registrierung/</cmdp:rights>
<cmdp:title xml:lang="de">Deutsches Referenzkorpus -
DeReKo</cmdp:title>
<cmdp:type>collection</cmdp:type>
```

Figure 1: VLO entry and excerpt of the underlying metadata file for a German corpus.

Some metadata schemas include basic support for information regarding the development process of a resource, e.g. the CMDI component *CreationToolInfo*⁷ and the *PreparationHeader* type at LAUDATIO's⁸ TEI customization.⁹ However, schemas to track processes are more common in disciplines where a fully automatic execution of a workflow is intended and thus tend to omit the possibility of manual steps such as inspection, extraction and annotation. In addition, traditional metadata is often collected at the end of the creation process, which is why we call it static metadata here.

An example of process metadata is generated in WebLicht¹⁰ (Hinrichs et al., 2010), where different web services can be selected to build a chain which acts on user-defined input. Figure 2 shows a chain consisting of a converter, a tokenizer and a part-of-speech tagger. Figure 2b shows XML-encoded process metadata, which is generated together with the chain and thus part of the output of the processing steps.¹¹

¹<http://dublincore.org/>

²<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html>

³<https://catalog.ldc.upenn.edu/>

⁴<http://www.resourcebook.eu/>

⁵<https://www.clarin.eu/content/component-metadata>

⁶<http://www.clarin.eu/vlo/>

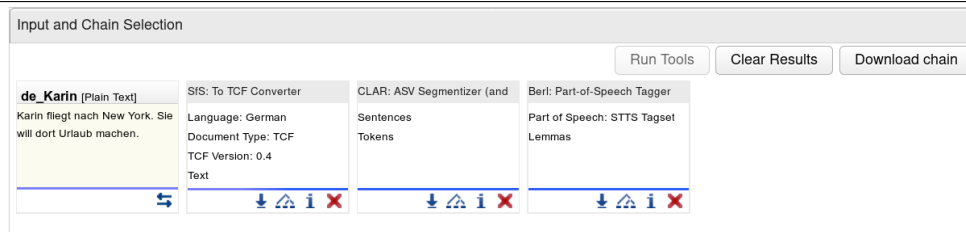
⁷Component [clarin.eu:cr1:c.1290431694497](https://catalog.clarin.eu/cr1:c.1290431694497) at <https://catalog.clarin.eu/ds/ComponentRegistry/>

⁸LAUDATIO-Repository, persistent identifier: <http://hdl.handle.net/11022/1007-0000-0000-8E65-F>

⁹https://korpling.github.io/LAUDATIO-Metadata/Preparation/teiODD-LAUDATIOPreparation_S7/document.html

¹⁰<https://weblicht.sfs.uni-tuebingen.de/weblicht/>

¹¹Process metadata of the chain without the results is also available from the 'Download chain' button in the upper left corner.



(a) Visual outline of the chain.

```
<cmd:Toolchain>
  <cmd:ToolInChain>
    <cmd:PID>http://hdl.handle.net/11858/00-1778-0000-0004-BA56-7</cmd:PID>
    <cmd:Parameter value="de" name="lang"/>
    <cmd:Parameter value="text/plain" name="type"/>
  </cmd:ToolInChain>
  <cmd:ToolInChain>
    <cmd:PID>http://hdl.handle.net/11022/0000-0000-94F4-5</cmd:PID>
  </cmd:ToolInChain>
  <cmd:ToolInChain>
    <cmd:PID>http://hdl.handle.net/11858/00-203C-0000-0024-7588-C</cmd:PID>
  </cmd:ToolInChain>
</cmd:Toolchain>
```

(b) Process metadata specifying the applied services and settings.

Figure 2: Example of a web service chain created with WebLicht.

3 The GRAIN Case

GRAIN is an annotated corpus based on German radio interviews (Schweitzer et al., 2018). It focuses on size of primary data and annotations and it contains various layers of annotation ranging from morpho-syntax, intonation and phonetic features to discourse level and document structure. Some annotations are created manually, some are created automatically. To keep track of the annotations available for a specific interview, as well as to document their development and interdependencies, a schema of process metadata is applied. The schema is based on the assumption that each step in the process of the corpus and annotation creation can be mapped to the simple set up of *input*, *operator* and *output*. Thereby operators can be automatic, e.g. tools for annotation, conversion or extraction, as well as manual, e.g. researchers creating or correcting annotations, choosing or deleting parts of the data, etc. Automatic operators are applied in a specific version and might also make use of input parameters or (exchangeable) components. For manual operators specific knowledge states, annotation parameters and applied guidelines, lexicons, etc. can be tracked. It is left to the author, if a file such as a trained model or a set of guidelines is part of the input or a component, such that several approaches can be reflected. Figure 3 shows the JSON-style ¹² schema applied with GRAIN.

```
{
  "result": [],
  "input": [],
  "workflowSteps": [
    {
      "description": "",
      "mode": "",
      "operators": [
        {
          "name": "",
          "version": "",
          "parameters": "",
          "components": [
            {
              "name": "",
              "version": "",
              "type": ""
            }
          ]
        }
      ]
    }
  ]
}
```

- (a) Basic structure.
- (b) Each entry can consist of several workflow steps, e.g. annotation and merging of annotations.
- (c) Each workflow step might comprise several operators, e.g. different annotators or different tools directly passing on their output.
- (d) Each operator might refer to several components, e.g. annotation guidelines and a lexicon.

Figure 3: Schema of process metadata in GRAIN.

¹²In true JSON (<http://www.json.org/>) arrays are ordered lists, however w.l.o.g. parallel or partly parallel work by annotators as well as an arbitrary component list can be captured here. An explicit marker for parallelism is not yet implemented.

4 The RePlay-DH Case

The RePlay-DH project (Gärtner et al., 2018) uses a process metadata schema very similar to the one described in the GRAIN section¹³. Its main focus is however on another often overlooked aspect of (process) metadata, namely its elicitation. It is quite common for scientific processes to be documented either ahead of time (e.g. in the form of a workflow plan which is then meant to be followed) or retroactively (in an attempt to describe afterwards what has been done). The potential for incomplete or incorrect documentation in both cases is bothersome: Minor changes to planned actions might be left out or seemingly unimportant details can be missed in the documentation. To address this issue a utility software (Gärtner et al., 2018) was developed to accompany researchers throughout their entire workflow and allow for flexible process documentation along the way. By using version control in the form of Git¹⁴ the RePlay-DH Client is able to (i) track fine-grained modifications, (ii) offer free navigation in the workflow history graph and (iii) store process metadata together with the actual physical changes¹⁵. While the default serialization format of the collected metadata is JSON¹⁶, the RePlay-DH Client also offers export options for other established formats such as P-PLAN¹⁷. This and the ability to customize the actual metadata schema makes it applicable for a wide range of workflows or research fields¹⁸. For detailed explanations of the metadata schema and the software itself we refer the reader to the papers mentioned above.

5 Conclusions

We discussed the value of metadata for the sustainability of resources, made a distinction between static and process metadata and showed examples for the collection and representation of process metadata. Tracking metadata while a resource is created documents the process for the creators and includes decisions which might be lost at a later stage. In cases where a resource cannot be made available, process metadata increases the possibility to estimate the comparability of studies conducted on or with them.

References

- Markus Gärtner, Uli Hahn, and Sibylle Hermann. 2018. Supporting sustainable process documentation. In Georg Rehm and Thierry Declerck, editors, *Language Technologies for the Challenges of the Digital Age*, pages 284–291, Cham. Springer International Publishing.
- Twan Goosen and Thomas Eckart. 2014. Virtual language observatory 3.0: What’s new? In *CLARIN Annual Conference*, Soesterberg, Netherlands.
- Markus Gärtner, Uli Hahn, and Sibylle Hermann. 2018. Preserving workflow reproducibility: The RePlay-DH client as a tool for process documentation. In Nicoletta Calzolari et al., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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.
- Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Roesiger, Antje Schweitzer, Sabrina Stehwiien, and Jonas Kuhn. 2018. German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection. In Nicoletta Calzolari et al., editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Dieter van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari et al., editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

¹³The two projects collaborated tightly and the annotation work in GRAIN actually doubled as a pilot project for the Git-based approach of collecting and storing process metadata as defined by RePlay-DH.

¹⁴<https://git-scm.com/>

¹⁵The tool shields users from the complexity of Git and can utilize commit messages to store process metadata in JSON.

¹⁶<http://www.json.org/>

¹⁷<http://www.opmw.org/model/p-plan/>

¹⁸For cases where the actual workflow is performed primarily in a local workspace.

The Best of Three Worlds: Mutual Enhancement of Corpora of Dramatic Texts (GerDraCor, German Text Archive, TextGrid Repository)

Frank Fischer
National Research University
Higher School of Economics, Moscow;
DARIAH-EU
ffischer@hse.ru

Susanne Haaf
BBAW; ZDL
haaf@bbaw.de

Marius Hug
BBAW; CLARIAH-DE
marius.hug@bbaw.de

Abstract

In most cases when tackling genre-related research questions, several corpora are available that comprise texts of the genre in question (like corpora of novels, plays, poems). This paper describes how to combine the strengths of different corpora to increase corpus sizes, correct mistakes and mutually enhance the depth and quality of the markup. The use case demonstrated regards three TEI-encoded corpora of German-language drama: the dedicated German Drama Corpus (GerDraCor) and the two implicit subcorpora of dramatic texts contained in the CLARIN-D-maintained German Text Archive (DTA) and the DARIAH-DE-run TextGrid Repository.

1 Introduction

In digital literary studies, research into the three traditional literary genres (poetry, prose, drama) is progressing on the basis of different types of corpora. The numerous collections of poems and novels are complemented by corpora of dramatic texts of various languages and sizes. Next to Paul Fièvre's collection of French classical drama, "Théâtre classique", comprising 1,290 plays to date, and "Shakespeare His Contemporaries", comprising 860 English plays, the German-language drama corpus GerDraCor is continuously growing to reach a similar order of magnitude, with 474 plays to date. The majority of plays was inherited and enhanced from the DARIAH-DE-run TextGrid Repository.

In this article we describe the mutual enhancement of three different corpora of dramatic texts: aforementioned GerDraCor, a corpus entirely dedicated to plays, as well as the two subcorpora of plays contained in the larger DTA and the TextGrid Repository. For this purpose, we set out to do three things:

1. enable the interoperability of texts from three corpora by following the recommendations of the DTA Base Format (DTABf) and implementing workflows to convert other TEI-based drama formats into DTABf;
2. mutually enrich and optimise all corpora in question;
3. ensure the long-term availability of the upgraded corpus data.

The integration of these three German-language drama corpora will happen in a controlled manner, meeting general infrastructural (e.g., DTABf as input format for DTA) as well as concrete research requirements (e.g., unequivocal speaker IDs). The integrated corpus is not only bigger, but also combines the different markups into a generally enhanced supercorpus of German-language drama.

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

Our work is to be seen in the context of different other projects bringing together digital corpora. To mention only two of them: 1. Early English Books Online (EEBO, <https://eebo.chadwyck.com>), to which more than 200 libraries contributed millions of page scans from works printed between 1473 and 1700, and 2. the Corpus of Literary Modernism (KOLIMO, <https://kolimo.uni-goettingen.de/>), a digital comparative corpus of German narrative literary Modernism, in which several already existing repositories (Gutenberg-DE and gutenberg.org as well as TextGridRep and DTA) got conjoined.

This paper also serves as pre-study for the integration of the DARIAH-DE-maintained “Digitale Bibliothek” within the TextGrid Repository and the German Text Archive (DTA), scheduled for the current CLARIAH-DE funding phase.

2 Differences in Markup

The three corpora under consideration have been enriched with TEI markup in three different projects. Consequently, the TEI dialects which were applied differ from each other, thus raising the issue of interoperability between texts. In order to bring together the three corpora, steps to harmonise the data have to be pursued. Examples 1 to 3 give insight into encoding variances by showing the same text passage from Goethe’s *Faust. Eine Tragödie* (1808) as it can be found in the three corpora.

2.1 Example 1: Deutsches Textarchiv

```
<milestone unit="section" rendition="#hr"/>
<pb n="[57]" facs="#f0063"/>
<milestone unit="section" rendition="#hr"/>
<div n="2">
  <head><hi rendition="#g">Vor dem Thor</hi>.</head><lb/>
  <milestone unit="section" rendition="#hr"/>
  <stage><hi rendition="#g">Spaziergãnger</hi> aller Art<lb/>
  ziehen hinaus.</stage><lb/>
  <sp who="#HANDWE">
    <speaker><hi rendition="#g">Einige Handwerksbur</hi>.
    </speaker><lb/>
    <p>Warum denn dort hinaus?</p>
  </sp><lb/>
  <sp who="#AND">
    <speaker><hi rendition="#g">Andre</hi>.</speaker><lb/>
    <p>Wir gehn hinaus auf</p>
  </sp><lb/>
  <sp who="#ERST">
    <speaker><hi rendition="#g">Die Er</hi>.
    <p>Wir aber wollen nach der Mu</p>
  </sp><lb/>
  [...]
</div>
```

2.2 Example 2: GerDraCor

```
<div type="scene">
  <head>Vor dem Tor.</head>
  <stage>Spaziergãnger aller Art ziehen hinaus.</stage>
  <sp who="#einige_handwerksburschen">
    <speaker>EINIGE HANDWERKSBURSCHEN.</speaker>
    <l>Warum denn dort hinaus?</l>
  </sp>
```

```

<sp who="#andre_handwerksburschen">
  <speaker>ANDRE.</speaker>
  <l>Wir gehn hinaus aufs Jägerhaus.</l>
</sp>
<sp who="#einige_handwerksburschen">
  <speaker>DIE ERSTEN.</speaker>
  <l>Wir aber wollen nach der Mühle wandern.</l>
</sp>
[...]
```

2.3 Example 3: TextGrid Repository

```

<div type="text" xml:id="tg1387.2">
  <div type="h4">
    <head type="h4" xml:id="tg1387.2.1">Vor dem Tor.</head>
    <stage rend="zenoPC" xml:id="tg1387.2.4">
      <hi rend="italic" xml:id="tg1387.2.4.1">Spaziergänger aller
      Art ziehen hinaus.</hi>
    </stage>
    <lb xml:id="tg1387.2.5"/>
    <sp>
      <speaker xml:id="tg1387.2.6.part1">EINIGE
      HANDWERKSBURSCHEN.</speaker>
      <l rend="zenoPLm4n4" xml:id="tg1387.2.7">Warum denn dort
      hinaus?</l>
    </sp>
    <sp>
      <speaker xml:id="tg1387.2.8">ANDRE.</speaker>
      <l rend="zenoPLm4n4" xml:id="tg1387.2.9">Wir gehn hinaus
      aufs Jägerhaus.</l>
    </sp>
    <sp>
      <speaker xml:id="tg1387.2.10.part1">DIE ERSTEN.</speaker>
      <l rend="zenoPLm4n4" xml:id="tg1387.2.11">Wir aber wollen
      nach der Mühle wandern.</l>
    </sp>
    [...]
  </div>
</div>
```

3 Consolidation of Speaker Information

The German Drama Corpus (GerDraCor, <https://draacor.org/ger>) as part of the Drama Corpora platform DraCor (Fischer et al. (2019)) inherited most of its current texts from “Digitale Bibliothek” (<https://textgrid.de/digitale-bibliothek>), which itself comprises TEI versions of the zeno.org text collection (<http://www.zeno.org/>). The DraCor platform delivers an API layer for all connected TEI corpora to facilitate research and make for an easier reproducibility. The modularity of the platform allows for a separation of corpus maintenance from the research-driven technology stack (Fig. 1).

Among other kinds of metrics, the API provides data for each play based on the co-occurrence of characters (speaking entities) in different scenes of a play. The data gets extracted from the underlying TEI at runtime and can be used for state-of-the-art network visualisations (via Gephi or similar programs). The social network extracted from Goethe’s *Faust. Eine Tragödie* may serve as an example (Fig. 2, left).

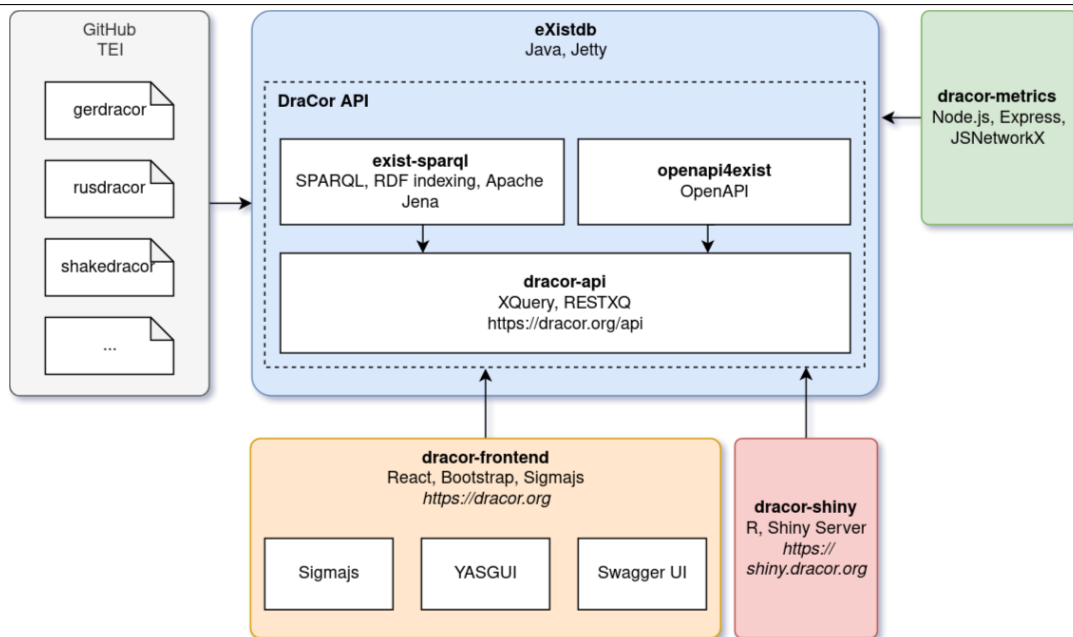


Figure 1: Technology stack of dracor.org: TEI corpora on the left.

The German Text Archive (DTA, <http://www.deutschestextarchiv.de/>) also features an implicit subcorpus of dramatic texts. The overlap between DTA and GerDraCor is 60, i. e., 60 of the texts are contained in both corpora. Since the origin of either corpus is fundamentally different and both involve different markup decisions, it is possible to mutually enhance both. While DTA’s drama collection is based on first editions and is more accurate thanks to a more sophisticated proofreading concept, GerDraCor has advantages when it comes to the identification of speaking entities.

As a consequence, extracting network data from the same *Faust. Eine Tragödie* in its DTA shape yields a different result (Fig. 2, right). The reason for this is the way IDs are automatically shortened and assigned within the DTA workflow. This operation sometimes results in ID overlaps between entities, which a human tagger would easily identify as different:

1. #IRR is assigned to “Irrlicht” in the Walpurgis Night scene as well as to the “Irrlichter” (plural!) in the later Walpurgis Night’s Dream (which is a different scene with completely different speaking entities);
2. #GEIST merges the “Geist” of the Night scene with the “Geist, der sich erst bildet” of the Walpurgis Night’s Dream.

As a result of incorrectly merged IDs, the formally unconnected Walpurgis Night’s Dream (bottom left in the DraCor graph in Fig. 2) is dragged into the main graph of the DTA-based visualisation (bottom right of the DTA graph in Fig. 2) via IDs #IRR and #GEIST.

For this and other reasons, the GerDraCor version of Goethe’s first part of *Faust* with 115 speaking entities is opposed by 104 such speaking instances in the DTA version. With regard to the thoroughly corrected IDs, the DTA version will benefit from GerDraCor data.

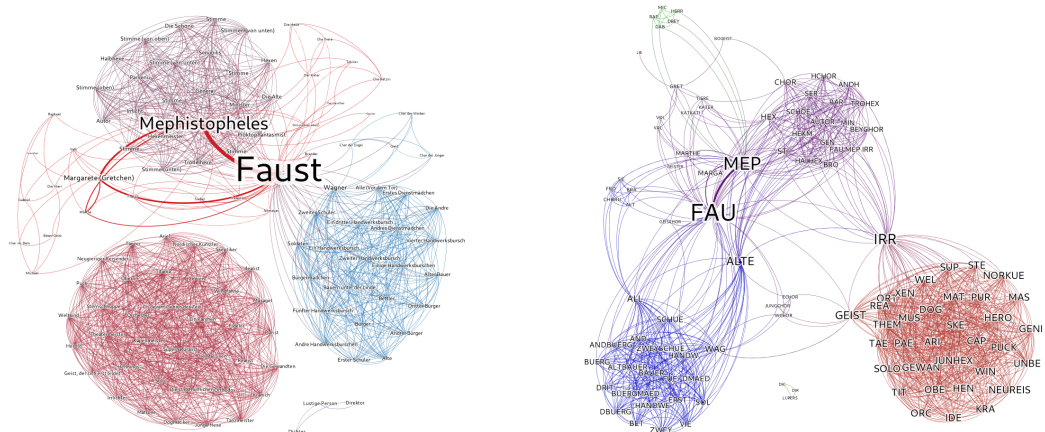


Figure 2: *Faust. Eine Tragödie*, co-occurrence graph extracted via DraCor's API (left) and from DTA (right), both graphs embellished with Gephi.

4 Consolidation of TEI Markup

The DTA corpus comprises a selection of currently 5,080 works of four genres (fiction, functional literature, scholarly works, newspapers) and around 130 text types, dating back to the 17th to 19th century (Geyken et al. (2018)). The corpus includes 91 dramas and another 10 works containing dramatic text. The DTA corpus is continuously growing: starting as a full-text digitisation project with a bibliography of 1,400 works to digitise from scratch, the DTA has by now evolved into a platform where digitised historical texts of high quality produced by different kinds of scholarly projects may be integrated and thus brought together after having been harmonised with the DTABf (see below). The goal is to aggregate scattered resources and to provide them as interoperable data for further analysis with tools provided by the DTA platform and beyond.

The strength of the DTA corpus and warrantor of interoperability is the well-established “DTA Basisformat” (DTABf) (Haaf et al. (2014)) which underpins the corpus. Interoperability issues affect different aspects, from metadata exchange through the extraction and analysis of document components up to the creation of a uniform stylesheet in order to present all corpus texts in a similar way (Geyken et al. (2012)). The DTABf forms a strict subset of the TEI tag set (i.e., no extensions of or content changes to the TEI P5 tag set are made) and is recommended by the German Research Foundation (DFG) as annotation or exchange format for editions and language corpora.

The DTABf has been developed to solve the problem of interoperability between TEI-annotated texts (Unsworth (2011)). The fact that similar source texts can take very different shapes in TEI complicates the exchange of TEI texts and the provision of project-independent tools for TEI data. Fig. 3 shows the usage of TEI elements for the above-mentioned *Faust. Eine Tragödie* in its GerDraCor, DTA and TextGrid Repository versions, exhibiting conspicuous differences in the interpretation and application of the TEI guidelines.

Adopting the DTABf for the German Drama Corpus (GerDraCor) which is mostly based on the TextGrid Repository will create a basis for mutual enhancements and ensure the interoperability of three corpora, combining the strengths of three worlds.

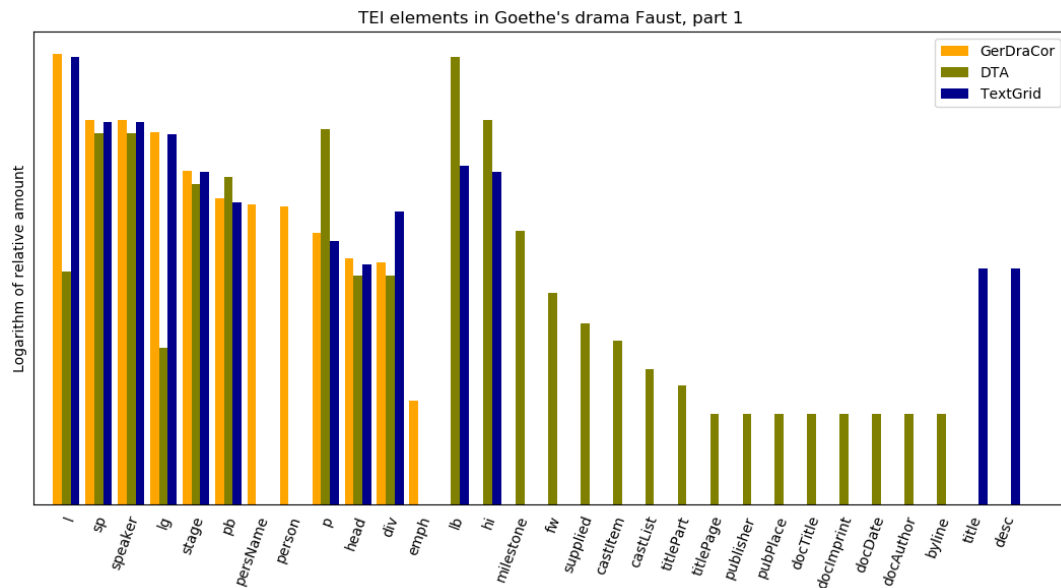


Figure 3: Usage of TEI elements in <text> for Goethe's *Faust. Eine Tragödie* (excluding <front>, <body>, <back>, including <person> and <persName> from <teiHeader> in GerDraCor).

5 Conclusion

In this talk we plan to present an aggregated corpus of German-language drama compiled from three different sources, processed and published according to the FAIR principles and available for free reuse by the scholarly community (CLARIN, DARIAH and beyond). We illustrate the possible heterogeneity of TEI-annotated texts by different examples and describe our work on TEI harmonisation in the context of dramatic texts based on DTABf as a recommended CLARIN input format. We present the new opportunities for corpus-based studies which evolve from the consolidation of the three resources. Finally, we want to evaluate our work on merging the drama corpora, its costs and benefits, with regard to the integration of resources from the TextGrid repository into the DTA as planned in CLARIAH-DE. The result is a richer, research-friendly corpus brought up to par with similar French and English corpora.

References

- Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtel, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor. In *DHd 2019. Digital Humanities: multi-medial & multimodal. Konferenzabstracts*, pages 194–197.
- Alexander Geyken, Susanne Haaf, and Frank Wiegand. 2012. The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora. In *11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing, Proceedings of the Conference (LThist 2012 Workshop)*, pages 383–391.
- Alexander Geyken, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In Henning Lobin, Roman Schneider, and Andreas Witt, editors, *Digitale Infrastrukturen für die germanistische Forschung*, pages 219–248. De Gruyter.

- Susanne Haaf, Alexander Geyken, and Frank Wiegand. 2014. The DTA “Base Format”: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. *Journal of the Text Encoding Initiative*, (Issue 8).
- John Unsworth. 2011. Computational Work with Very Large Text Collections. *Journal of the Text Encoding Initiative*, (Issue 1).

Mapping METS and Dublin Core to CMDI: Making Textbooks Available in the VLO of CLARIN

Francesca Fallucchi

Innovation and Information Engineering Dep.,
G. Marconi University, Rome, Italy
Digital Information and Research
Infrastructures Dep., Georg Eckert Institute for
International Textbook Research,
Braunschweig, Germany
f.fallucchi@unimarconi.it
fallucchi@leibniz-gei.de

Ernesto William De Luca

Innovation and Information Engineering
Dep., G. Marconi University, Rome, Italy
Digital Information and Research
Infrastructures Department, Georg Eckert
Institute for International Textbook Research,
Braunschweig, Germany
ew.deluca@unimarconi.it
deluca@leibniz-gei.de

Abstract

In a time where the amount of digital resources and the complexity of the relations between them are expanding rapidly and unpredictably it is necessary to manage and to find electronic resources. Descriptive metadata characterise a resource with keyword-value pairs. The use of such descriptions allows researchers clearer and easier access to available resources. In this way, users can manage and find research data beyond traditional publications. The Georg Eckert Institute (GEI) creates and curates various digital resources that are offered to the community of international textbook research and other scientific field. This paper discusses how to provide a CMDI (Component MetaData Infrastructure) profile for our textbooks, in order to integrate them into CLARIN infrastructure and thus open them make them fairer. After adapting to CMDI profile for research project data, we now look into the creation of a new profile for the digitized historical textbooks of "GEI-Digital". We describe our workflows and the adversities and problems faced when trying to convert METS metadata into CMDI.

1 Introduction

The systematic description of books, sound and video recordings, and other artefacts, has a long tradition in the field of LIS (Zinn et al. 2016). A number of most used metadata standards to record resource information in the context of libraries are MARC, Dublin Core (DC), METS, etc. (Hillmann et al. 2008); (DCMI 2018). Metadata are stored and have to be managed carefully remaining a crucial aspect in the life cycle of language resources, especially, if we analyse the schemas, structural standards for metadata within the fields or elements where the data resides and new developments in metadata infrastructures, such as CMDI provided by CLARIN¹. In (Fallucchi and De Luca 2019) we explain the connection and mapping of knowledge representations between RDF and CMDI.

Hereby, we present a digital infrastructure of our textbook-related services and data for GEI², which are available and open for researchers worldwide. We started a process to established standards and provide APIs for other services, in order to integrate our resources and tools into the CLARIN infrastructure and make them discoverable in the VLO (Virtual Language Observatory). The integration makes our data more compatible, visible, open for other communities. In particular, our focus is to use CMDI as a unique metadata descriptive standard. At the GEI, we currently use it for two sets of data. Firstly, for managing the TEI-headers of the digital Edition "*WorldViews*"³, a project that publishes sections of textbooks as well as translations, essays and additional data on authors and publishers. Secondly, we would to use CMDI for "*GEI-Digital*"⁴, a large corpus of digitized historical textbooks currently containing more than 6200 books with more than a 1,5 million pages. In (Fallucchi et al. 2019) we

¹ <https://www.clarin.eu/>

² <http://www.gei.de/en/home.html>

³ worldviews.gei.de/

⁴ <http://bibliothek.gei.de/en/gei-digital.html>

described a process to build resources using a CMDI description and we tested it on the TEI headers resources from the WorldViews digital repository of the GEI.

In this paper, we extend our idea and focus on the creation process for a CMDI metadata profile which fulfills the needs of GEI-Digital project. GEI-Digital can be exported, via an OAI-PMH-Interface using the resource metadata within the following standard: METS, MARCXML, DC, ESE and OLAC. The library likes and uses METS because they want to record descriptive, administrative, and structural metadata for a digital library object. The rest of the paper is organized in the following: in Section 2 we describe the mapping study from our METS resources to CMDI and we propose a draft of METS Profile (Section 2.1). In Section 2.2 we describe the mapping problems that leads to the decision to stop the CMDification process and we introduce the current solution found to making textbooks available in the VLO of CLARIN. In Section 3 we give some conclusions and an outlook on future works.

2 METS to CMDI feasibility study

The first important step in the creation of CMDI records was to establish a conceptual model of the resources. One of the most important processes related to the planned formalisation is the digitisation and metadata editing process for resources. METS is considered the most complete standard that provides a means of encoding digital library materials for GEI-Digital repository⁵. It is created automatically by the Goobi software⁶ which is used in the digitalization workflow and provides a fine granularity, if somewhat fuzzy, standard for encoding administrative and structural metadata regarding objects within a digital library. The software organises the workflow into separate and clearly distinguishable steps. These steps are organised into templates. Each template combines all necessary steps into a structured workflow that allows all relevant data to be created, input and exported for the presentation of each resource type in the GEI-Digital project. By analysing the Goobi output of the GEI-Digital workflow and participating in a number of discussions between librarian, domain experts, information scientist experts and computer science technicians we have been able to define the conceptual model necessary to model the existing resources in the CMDI world.

2.1 METS CMDI Profile

In CMDI the syntax of a metadata set for a given resource type can be defined with reference to the semantics of each element in the CLARIN Concept Registry (CCR)⁷. To ensure semantic interoperability, the component elements have to be linked to data categories in the Data Category Registry (DCR)⁸. In the CMDI Component Registry⁹, we can either reuse existing CMDI components and profiles (as we did in case of WorldViews) or define new ones, and therefore either map to existing structures and content, or model the structure of the metadata set from scratch. Unfortunately, a profile that can be reused to describe GEI-Digital METS resources does not exist in the CMDI Component Registry. There is only a generic *Book* profile ID (ID: clarin.eu:cr1:p_1345561703682) in CMDI Component Registry and as written in its description it is a *SAMPLE PROFILE! It has not been designed for publication purpose or any other specific use other than demonstrating part of the functionality of CMDI* and has not relevant information to describe bibliographic data of our textbooks. In fact this profile is not adequate for the bibliographic data that we have for the textbooks of GEI-Digital repository. So, we try to create a new METS Profile in CMDI Component Registry in accord with the METS document structure of our resources. For the traditional metadata standard METS, the respective “METS_CMDI Profile” should provide a standardized XML format for transmission of complex digital library objects between systems. To achieve this semantic interoperability, we create the “METS-CMDI” profile and reference each element in CMDI Component Registry that establishes the container for the information being stored and/or transmitted by the standard. We now briefly describe the major components and elements of the profile. The profile consists of the CMDI components, which are: *METS Header*, *dmdSec*, *amdSec*, *fileSec*, *structMap*, *structLink*, and *behaviorSec*. In Fig. 1. we show the METS Profile, its attributes, elements and components that we created in *GEI workspace* of CMDI Component

⁵ <http://gei-digital.gei.de/viewer/>

⁶ <https://www.intranda.com/en/digiverso/goobi/goobi-overview/>

⁷ <https://concepts.clarin.eu/ccr/browser/>

⁸ <https://www.iso.org/sites/dcr-redirect/dcr.html>

⁹ https://catalog.clarin.eu/ds/ComponentRegistry/#/?_k=jrrech0

Registry as a result of the mapping process started from the bibliographic data elements METS of our resources.

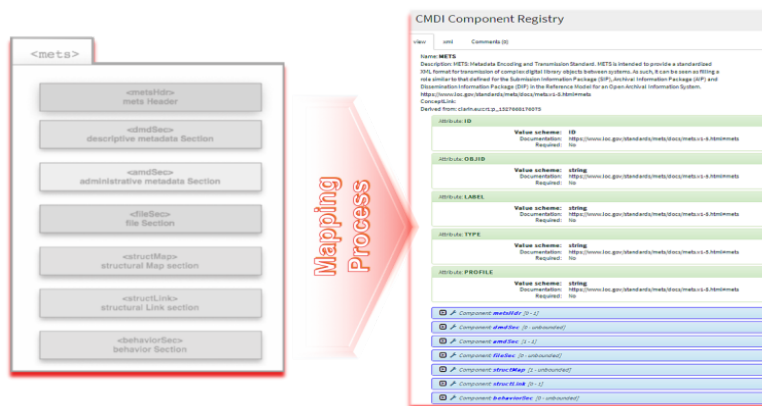


Fig. 1 METS CMDI Profile

The *METS Header* should record metadata about the METS document itself (not the digital library object that the METS document encodes). It has two possible subsidiary elements: agent (document agent) and altRecordID. To map this contents we created in CMDI Component Registry the the CMDI component *METS Header* with 5 attributes: *ID*, *ADMIN*, *CREATEDATA*, *LASTMODIFIED* and *RECORDSTATUS*, an elements altRecordID and an component agent. Agent is a component because *METS Header* could have not only attributes but also elements. In the Fig. 2 we report the steps for mapping process how drove the METS Header component creation in CMDI Component Registry which will host bibliographic data of our GEI-Digital repository. The figure highlights how the mapping was carried out starting from the analysis of the METS bibliographic data and show the match between the metadata in the XML file and elements, attributes and components CMDI creating in CMDI Component Registry.



Fig. 2 Map of bibliographic data to METS Header component

The similar approach is used to create the other CMDI components. The *dmdSec* Component allows describing the all of the descriptive metadata for all items in the METS object. The *amdSec* Component records all of the administrative metadata for all items in the METS object. The *fileSec* Component allows to record information regarding all of the files which comprise the digital library object. The *structLink* Component allows to describe the specification of hyperlinks between different components. The *behaviorSec* Component enables one to describe the information about a service that is associated with the METS object.

2.2 Mapping problems

Many of the profile's components, elements and their possible values have a semantic definition by a link to an entry in the CMDI Component Registry. For the ones that were lacking we created definitions and provided other relevant information required for inclusion into the CCR. We submitted the complete table of missing elements, in the format required, to the maintainers of the CCR but CMDI mapping problems led the CLARIN metadata managers, to stop the CMDification METS process. During our mapping work with METS, the following issues came up/occured:

- CMDI cannot manage recursive structures like those based on the `<mets:div>` components which are mandatory to map the physical layer of the digital objects. METS semantics states that: "The structural divisions of the hierarchical organization provided by a `<structMap>` are represented by division `<div>` elements, which can be nested to any depth."
- CMDI cannot manage the cross-reference links like `<mets:smLink>` components, which can be found in any METS structure.
- CMDI at the moment does not provide any semantics in its own Concept Registry for mapping a METS structure.
- CMDI cannot manage different concepts tied to the same label. In METS such cases are disambiguated by their position into the XML structure.

The METS to CMDI mapping revealed itself impracticable. Extracting and mapping only the MODS sections from the whole METS structure is a vain effort because of its intrinsic limitations. Therefore, after considering which metadata standards are currently supported by the CLARIN Infrastructure, we planned to map our GEI-digital metadata from the DC format to CMDI using our OAI-PMH server.

3 Conclusion

In this paper, we describe a mapping process to established standards and provide APIs for other services, in order to integrate our resources and tools into the CLARIN infrastructure and make them discoverable in the VLO. We focus on CMDI as a unique metadata descriptive standard for encoding administrative and structural metadata of the resources of GEI-Digital repository. To achieve this semantic interoperability, we create the METS CMDI profile but the CMDification process is stopped because for mapping problems, the METS to CMDI mapping revealed itself impracticable. The CLARIN metadata managers, which supported us in including our data in the VLO, stated in one coordination meeting that using Dublin Core is an acceptable approach although a bit of richness and semantics may be lost compared to the METS/MODS metadata. In spite of the classical DC, CMDI has an OLAC-CMDI mapping ready to be used, (as suggested by the CLARIN Infrastructure Managers), which can be adopted to manage DC metadata. GEI-Digital metadata from the DC is just an intermediate solution but does not satisfy our library team. This solution has temporarily avoided blocking the activities for making textbooks available for the harvesting in the VLO. Our library team likes and used METS because they want to record more than is possible with DC. We tried this mapping process from METS to CMDI and this and that did not work because of CMDI cannot manage recursive structures and the cross-reference. We are in the process of deciding what to do. We would like to find a more richness and relevant profile that satisfy our library team because they not be happy of the richness and semantics of the DC mapping.

Reference

- DCMI: DCMI Metadata Terms. 2018. Available: <http://www.dublincore.org/documents/dcmi-terms/>.
- F. Fallucchi and E. W. De Luca. 2019. Connecting and Mapping LOD and CMDI Through Knowledge Organization, Springer, Cham, 2019, pp. 291–301.
- F. Fallucchi, H. Steffen, and E. W. De Luca. 2019. Creating CMDI-Profiles for Textbook Resources,” Springer, Cham, 2019, pp. 302–314.
- D. I. Hillmann, R. Marker, and C. Brady. 2008 Metadata standards and applications, Ser. Libr., vol. 54, no. 1–2, pp. 7–21, 2008.
- C. Zinn, T. Trippel, S. Kaminski, and E. Dima, 2016 Crosswalking from CMDI to Dublin Core and MARC 21, Int. Conf. Lang. Resour. Eval., no. i, pp. 2489–2495, 2016.

What got connected - PARTHENOS ending

Matej Durco ACDH-OEAW Vienna, Austria matej.durco@oeaw.ac.at	Klaus Illmayer ACDH-OEAW Vienna, Austria klaus.illmayer_	Stefan Resch ACDH-OEAW Vienna, Austria stefan.resch_
--	--	--

Abstract

The RI-cluster project PARTHENOS is coming to an end after four years of intensive work. One of the main goals was the integration of metadata from the diverse domains represented by the partners. To this end, a common semantic model has been devised, aimed to capture the main entities of the knowledge generation process as they are present in resource metadata. In this paper we elaborate on the results of the aggregation process with the (traditional) focus on metadata quality.

1 Introduction

PARTHENOS (Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies)¹ is a project funded by the European Commission Horizon 2020 framework programme that started in May 2015 running for four years. The project empowers digital research in the fields of history, language studies, cultural heritage, archaeology, and related fields across the (digital) humanities. It aimed to establish interoperability between several existing research infrastructures, allowing to find, use and combine information from different domains. Consequently a central endeavour was the harmonisation and aggregation of heterogeneous data coming from the participating research infrastructures into a common semantic framework called PARTHENOS Entities Model (PEM) (Bruseker et al., 2018) based on CIDOC-CRM². CLARIN was a major partner in PARTHENOS project with regard to language resources and language studies in general. It operates the biggest catalogue of language resources in Europe, Virtual Language Observatory (VLO), since 2010. This catalogue aggregates the metadata about the resources from over 60 data providers, containing more than a million records. The backbone of CLARIN and the VLO is CMDI (Component Metadata Infrastructure) (Broeder et al., 2011). After we described the work on mapping CMDI to the PE model in (Durco et al., 2018), in this paper we present the preliminary results of the project with respect to metadata harmonisation and aggregation.

2 Mapping & Aggregation

PARTHENOS project set out with the promise of harmonizing metadata schemas from different sources and contexts in one common semantic framework. In general, the project delivered on its promise. Representatives of participating infrastructures and initiatives defined mappings from the local XML-based schemas to the PARTHENOS Entities Model using the 3M tool (Minadakis et al., 2015). These mappings have been applied on the respective metadata sets and the resulting RDF has been made available via a triple store. It has been also further converted and ingested into different dissemination endpoints, especially the CKAN-based³ Joint Resource Registry (Aloia et al., 2017), the central catalogue of the PARTHENOS data space. The mappings are expressed in X3ML mapping definition language⁴. The

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

¹<http://PARTHENOS-project.eu>

²<http://cidoc-crm.org/>

³<https://ckan.org/>

⁴<https://github.com/delving/x3ml/blob/master/docs/x3ml-language.md>

final mappings are published via zenodo⁵. These mappings serve as input for the customisable aggregation infrastructure, D-Net (Manghi et al., 2014), which powers a number of aggregation initiatives, among others the large-scale research publication portal OpenAire⁶. Both applications, D-Net and 3M, have been integrated into the central infrastructure, d4science, run by IST, Pisa powered by the software solution gCube⁷, used as the central provisioning platform for integrating content and services of PARTHENOS.

The common model PARTHENOS Entities Model is by design lossy, i.e. it does not try to capture all characteristics for the heterogeneous set of resources as they can be found in the local schemas, but rather concentrates on main entities and relations between them. These main entities are: Actor, Dataset, Service, Software, and Place.

2.1 CMDI Mapping

While other participating infrastructures could concentrate on manually crafting the mappings from their source XML-schema to the target model, in the case of CMDI, with its multitude of schemas a manual mapping approach was not feasible. Therefore as described in (Durco et al., 2018), we developed a dedicated application which generates mappings for individual CMDI schemas or profiles based on a template in combination with the interoperability mechanisms built into CMDI (Broeder et al., 2011). These mappings were made available to the aggregation workflow by publishing them on github⁸. The set of CMDI records as it is provided by the CLARIN VLO harvester⁹ had to be adjusted as well, regrouping the metadata records by profiles rather than the source collections. Currently still only records from six profiles are aggregated, even though mappings exist for all the profiles, due to capacity constraints with the responsible project partner.

3 Statistics

In the iterative process of mapping, evaluating and aggregating the data, the project team amassed a large set of descriptive statistics about the number of occurrences of various phenomena in the dataset. These are indispensable for getting an overview and for understanding the dataset, as well as input for any quality assessment or curation work. In the following we present a brief glance at this information¹⁰.

The overall dataset contains 46.7 Mio. triples with 8.6 Mio. distinct subjects of 75 distinct classes, described with 117 distinct properties. Table 1 gives account on the number of instances for selected main classes (including also the subclasses) from two points in time, to give an idea of the dynamics of the dataset.

Class	# instances (2019-01)	# instances (2019-04)
E39 Actor	172 956	265 758
PE18 Dataset	497 182	668 582
PE1 Service	389 541	507 620
D14 Software	26	31
E53 Place	47 591	144 625

Table 1: Number of instances for main classes (including subclasses)

Another important dimension is that of the metadata sources. The provenance is encoded via named graphs, keeping track of the aggregation source for each entity and each triple. Table 2 lists the number of triples per contributing infrastructure. These numbers are subject to numerous factors and may only be interpreted as rough indication of the size of the contributed dataset. PARTHENOS as provider refers to

⁵<https://doi.org/10.5281/zenodo.2574524>

⁶<http://openaire.eu/>

⁷<https://www.gcube-system.org/>

⁸http://github.com/acdh-oeaw/PARTHENOS_mapping

⁹<https://vlo.clarin.eu/oai-harvest-viewer/>

¹⁰Due to the continuous aggregation process, the dataspace is in flux. The presented numbers are prevalently from April 2019

Provider	# triples	Provider	# triples
ARIADNE	16 496 824	LRE MAP	442 946
CLARIN	11 932 600	Cultura Italia	408 237
CENDARI	9 911 824	METASHARE	246 707
Huma-Num - Nakala	3 716 181	DARIAH-GR	43 320
EHRI	1 495 762	PARTHENOS	7 843

Table 2: Number of triples per provider

a small hand-crafted set of relevant entities which however are not described in any other of the sources, e.g. CLARIN VLO as a service or CLARIN-ERIC as Research Infrastructure.

4 Metadata quality issues

Critical issue in a large-scale heterogeneous aggregation endeavour as pursued in PARTHENOS is the quality of the (meta)data. Unsurprisingly, we have encountered the usual classes of problems with data quality in aggregation scenarios. All listed issues have influence on the quality of the resulting harmonized metadata and dramatically hamper the discoverability of resources and the overall user experience:

- **Missing data** – There are often large lacunae in the aggregated data space. This can be due to erroneous or incomplete mapping, but mostly it is already the source metadata that does not make certain characteristics of a resource explicit.
- **Missing Labels** – While the mapping tool allows and encourages the definition of labels (human readable strings denoting an entity) for all generated entities, there is still a substantial portion of entities without a label (5,729,903 out of 8,581,496). To some extent this can be attributed to auxiliary entities dictated by the data model, not directly relevant for the user, and thus not needing a descriptive label. However even for the main entity types the coverage is suboptimal: around 60% for Dataset or Place, 33% for Actor.
- **Literals referring to entities** – Ideally a reference to an entity should be done with an unambiguous identifier, a URI. However, oftentimes, due to limitations of the source metadata schema and/or the metadata authoring tools, simple literals are used to denote entities like persons or institutions. This approach is inherently prone to spelling variations and ambiguous references. PEM offers a well-defined way to model these entities, however the problems in the source data counteract this potential. A major challenge in the mapping effort was to generate PEM entities out of this underspecified references, especially to generate a sensible, stable URI denoting given entity.
 Due to the principal uncertainty when trying to disambiguate a literal reference to an entity, a policy has been adopted, that if the same literal is encountered in the context of one provider it is considered the same identical entity. However when identical literals come from different sources, the probability that they don't refer to the same entity is deemed higher, therefore in such case two distinct entities are created, even though with identical label or appellation. The merging of identical entities is left for a separate curation step.
- **Variability of descriptors** - related to the previous problem, values in fields oftentimes come in various spellings or language variants, leading to generation of separate entities.
- **Underspecified semantics in the original metadata schema** – When mapping the source schemas to PEM, the meaning of certain elements in the source schema is sometimes not well defined. Typical example is the convolution of instances of Service and Software. While in the source formats as in colloquial use these are oftentimes used interchangeably, in PEM these are semantically clearly separated, Software as a subclass of a Digital Object, Service as subclass of an Activity.

- **Places & Spatial Coordinates** – For many instances of *E53.Place* geocoding information is missing or formatted in a non-standard way. Out of over 47.000 Places (as of 2019-01), only around 6.100 come with geocoding information, all from a single provider (CulturaItalia).

Places would lend themselves ideally for post-aggregation enrichment, automatically matching the instances against large-scale reference resources like wikidata or geonames.

- **Image representation** – Also in case of images (e.g. as photo of the described artefact) the data space is quite scarce, all available data (around 5,000 images) coming again just from one provider (CulturaItalia).
- **Same information in many graphs** – Presumably due to modelling error in the mapping process certain triples are repeated many times. E.g. the type of a collection consisting of many individual datasets may be indicated repeatedly in the context of every dataset. Albeit this is technically not incorrect, it clutters the data space with duplicate information leading to confusion of the user and potential problems (unexpected expansion of the result) in complex queries. Top 100 most often reoccurring triples appear in the dataset over 3 mio. times.

One last finding, not directly related to metadata quality, was the frustrating experience of having to deal with the inherent complexity of CMDI. While the other infrastructures could manually create mappings for their respective schema, in CMDI case a separate processing step had to be introduced, where the mappings have been generated based on the individual CMD profiles. Also substantial effort was required to explain the intricacies of CMDI to the project partners, as well as in the mapping and aggregation process, where additional measures had to be introduced for processing the CMDI records.

5 Conclusion

This abstract gave a glimpse on the results of the metadata aggregation and harmonisation effort in the project PARTHENOS, focusing on the descriptive statistics and metadata quality issues. Given the space constraints, the evaluation is rather superficial and will be elaborated upon further in the full paper.

Among benefits for CLARIN is an extensive experience with integrating its own ecosystem with that of related infrastructures, and – in the realm of metadata – groundwork for possible sound ontological grounding for CMDI. Last but not least, the mapping and aggregation process in PARTHENOS exposed yet again the high cost of the flexibility of CMDI.

References

- Nicola Aloia, Leonardo Candela, Franca Debole, Luca Frosini, Matteo Lorenzini, and Pasquale Pagano. 2017. PARTHENOS D5.2 Design of the Joint Resource Registry, April.
- Daan Broeder, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck, and Andreas Witt. 2011. A pragmatic approach to xml interoperabilitythe component metadata infrastructure (cmdi). In *Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies*, volume 7.
- George Bruseker, Martin Doerr, and Maria Theodoridou. 2018. PARTHENOS D5.5 Report on the Common Semantic Framework, October.
- Matej Durco, Matteo Lorenzini, and Go Sugimoto. 2018. Something will be connected-semantic mapping from cmdi to parthenos entities. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, number 147, pages 25–35. Linköping University Electronic Press.
- Paolo Manghi, Michele Artini, Claudio Atzori, Alessia Bardi, Andrea Mannocci, Sandro La Bruzzo, Leonardo Candela, Donatella Castelli, and Pasquale Pagano. 2014. The d-net software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, 48(4):322–354.
- Nikos Minadakis, Yannis Marketakis, Haridimos Kondylakis, Giorgos Flouris, Maria Theodoridou, Gerald de Jong, and Martin Doerr. 2015. X3ml framework: An effective suite for supporting data mappings. In *EMF-CRM@ TPD*, pages 1–12.

A New Gold Standard for Swedish NERC

Lars Ahrenberg

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

Leif-Jöran Olsson

Department of Swedish
University of Gothenburg, Sweden
leif-joran.olsson@svenska.gu.se

Johan Frid

Lund University Humanities Lab
Lund University, Sweden
johan.frid@humlab.lu.se

Abstract

Starting in 2018 Swe-Clarín members are working cross-institutionally on special themes. In this paper we report ongoing work in a project aimed at the creation of a new gold standard for Swedish Named-Entity Recognition and Categorisation. In contrast to previous efforts the new resource will contain data from both social media and edited text. The resource will be made freely available through SpråkbankenText.

1 Introduction

Named-Entity Recognition and Categorisation, henceforth NERC, is a standard task in NLP which has great value for many applications, including research in the humanities and social sciences. While many new methods are developed, there is a lack of data for training and testing. For Swedish, the most recent NERC gold standard is made from the Stockholm-Umeå Corpus, SUC, (Gustafson-Capková and Hartmann, 2006), which reflects the language of the late 20th century. This means that there are no examples reflecting linguistic usage in social media.

For this reason a collective effort to develop a gold standard NERC resource for Swedish has been started within the Swe-Clarín consortium. This paper presents the results of the project as they are in August 2019. Here we give an overview of the motivation and aims, the principles for annotation and some initial results.

2 Related work

NERC established itself as a standard task in NLP in the context of information extraction (Nadeau and Sekine, 2007). With the advent of social media and NLP being applied to user-generated texts it has met with new challenges (Derczynski et al., 2015). While earlier systems relied to a large extent on pattern matching and gazetteers, which worked well on edited text, the variation found in user-generated text and the output from OCR-systems, machine learning, and deep-learning approaches in particular, are becoming more and more powerful (Yadav and Bethard, 2018).

The first larger gold standard for named entities in Swedish text was the Stockholm-Umeå corpus (SUC), which was supplied with named-entity annotation for its second version (Gustafson-Capková and Hartmann, 2006). In the most recent version, SUC3.0, the annotation has been checked further. The NE annotation uses 10 categories, one of which is Other. The distribution is uneven over the categories with Numerical having the most (18098), and Events the fewest (245).

A joint Nordic project developed several systems for NERC on Scandinavian languages using six categories: PRS (Person), LOC (Location), ORG (Organization), EVT (Event), WRK (Work of Art) and OTH (Other) (Johannessen et al., 2005). The project compared several methods, both small-scale by

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

evaluating output manually, and on larger gold-standards. A conclusion of the project was the importance of gazetteers for achieving a good performance.

A Swedish system taking part in the project was SweNER (Kokkinakis, 2004). It has a layered architecture using large lists of known multiword names and ordered patterns and rules. It also uses a hierarchical taxonomy of categories. This system was later converted and re-implemented with the Helsinki Finite-State Transducer Technology, HSFT (Kokkinakis et al., 2014). The latter system had eight categories: Person, Location, Organisation, Artifact, Work&Art, Event, Measure/Numerical, Temporal. The evaluation was done with SUC3.0 as the gold standard, with some complications in matching the system's eight categories to the gold standard's ten.

(Ek et al., 2011) developed a NERC-system for short text messages. They used five categories, persons, locations, dates, clock times, and telephone numbers. As part of the project a corpus was collected comprising some 60,000 tokens which was annotated for the five categories.

3 Aims and process

As our project has a limited budget it was clear from the beginning that it should have a restricted scope. However, it should be expandable and possible to continue in the future. For this reason much effort has been given to the annotation guidelines and the analysis of critical examples. It was also decided to use a simple non-hierarchical taxonomy where each category is delineated both semantically and syntactically. This also means that certain types of entities that arguably are referred to by names or standard expressions, such as measures, are not covered.

In addition to the three most common categories, Person, Location, and Organisation¹, we included temporal entities and two *general* categories that are notoriously varied and difficult to recognize such as Event and Works Of Art (including Product names). In addition, two categories from the medical domain, another important domain for NERC systems, were included. The choice of the categories Symptom and Treatment is based on the i2b2/VA challenge (Uzuner et al., 2011) where Medical problem—treatment relations were one of the major tasks. Interest in clinical text mining is rising, but these texts are usually very hard to come by (for privacy reasons) so the inclusion of medically related categories in this project essentially means that we provide what may be the first publicly available medically related annotations of any kind for Swedish. The selected categories are shown in Table 1. Our aim is to reach at least 1000 instances for each category.

Category	Label	Category	Label
Person	PRS	Temporal	TME
Location	LOC	WorkOfArt/Product	WRK
Organisation	GRO	Symptom	SMP
Event	EVN	Treatment	MNT
Non-NE Token	O		

Table 1: The eight categories used in the NERC resource.

To simplify further it was decided that a token can only be annotated by one category. In case a token is eligible for two categories, the one applying to the longest expression is used. Thus, all tokens of a composite name such as *Lund University* will be tagged GRO, for organisation, although Lund is a town that could be tagged LOC when appearing on its own. Tokens that are not part of a name are tagged O.

The corpus has been compiled from the resources of the Swedish Language Bank using social media texts from blogs and discussion fora and edited text from newspapers. Recently texts from popular science, including Wikipedia, have been added. Some texts have been scrambled for copyright reasons, whereas for others the sentences occur in their natural order.

¹GRO rather than ORG is used as an abbreviation for Organisation in the annotation process to avoid inadvertent connotations with O.

Category	Iter 1	Iter 2	Iter 3
Person	0.898	0.946	0.930
Location	0.890	0.889	0.917
Organisation	0.759	0.789	0.846
Event	0.646	0.724	0.787
Temporal	0.463	0.699	0.836
WorkOfArt/Product	0.780	0.763	0.822
Symptom	0.537	0.664	0.750
Treatment	0.463	0.699	0.836
All	0.716	0.800	0.856

Table 2: Annotation progress. Fleiss' kappa at different stages in the annotation process.

3.1 Work process and annotation guidelines

A first collection of guidelines was compiled from earlier projects. An initial selection of texts were then annotated by three annotators. Inter-annotator agreement was not very high, in part depending on a number of problematic examples where the guidelines were insufficiently specified. After discussion, the guidelines were extended and supplied with more examples and a new round of annotation followed by the same annotators. While agreement on the first texts had now increased agreement on added texts were not at the same level. Thus, the process was iterated once more. The progress in annotation and inter-annotator agreement is shown in Table 2.

The project is recruiting volunteer annotators among Swe-Clarín partners to reach the goal of a minimum of 1000 instances/category. Although this is likely to initially reduce inter-annotator agreement, it will in the end improve the quality of the resource and increase the chances that it will be used and extended. The guidelines are still work in progress, but will be published with the resource.

4 Discussion

As can be seen in Table 2 inter-annotator agreement is steadily increasing. Most differences concern the question whether a token should be labelled as part of a named entity or not, i.e., O versus the other labels (see Table 3). This can sometimes be blamed on inattentiveness from one or other of the annotators, but also shows the need for clearer guidelines. Two of the categories have reached a value above 0.9 in Fleiss' kappa, which we judge as satisfactory. Conversely, another two are still below 0.8, thus below what (Landis and Koch, 1977) considers the threshold for "almost perfect agreement". One category with low agreement is Event, which also is a category with very few mentions in the data annotated so far. This also means that there are not so many examples in the data that can be taken as a basis for strengthening the guidelines. In particular, we have a number of instances of TV-shows, where annotators have problems distinguishing Event from WorkOfArt/Product. The other category with low agreement is Symptom, where a major problem for annotators has been to distinguish medical problems from normal variations in a person's well-being ('very tired', 'headache'), which, according to the guidelines, should not be annotated. Currently, the corpus consists of about 92,000 tokens annotated at least once, of which 92 percent are O. Table 3 also shows that the categories are unevenly distributed in the data annotated so far. For those with fewest instances, Event and Treatment, we have only reached about 5% of what we want. For this reason other sources such as popular science have now been included.

A2	PRS	LOC	GRO	EVN	TME	WRK	SMP	MNT	O
A1									
PRS	572	0	1	0	0	2	0	0	8
LOC	0	183	7	0	0	9	0	0	7
GRO	4	4	272	0	0	2	0	0	18
EVN	0	0	1	44	0	0	0	0	3
TME	0	0	0	0	535	0	0	0	42
WRK	2	0	5	1	3	475	0	0	56
SMP	0	0	0	0	0	0	222	0	52
MNT	0	0	0	0	0	1	1	53	21
O	7	0	12	3	65	66	28	10	37100
A3									
A1									
PRS	539	2	17	0	0	13	0	0	12
LOC	0	196	6	0	0	0	0	0	4
GRO	2	1	274	0	0	19	0	0	4
EVN	0	0	0	40	0	5	0	0	3
TME	0	0	0	0	536	0	0	0	41
WRK	1	1	6	0	0	502	0	0	32
SMP	0	0	0	0	0	0	262	0	12
MNT	0	0	0	0	0	0	3	64	9
O	58	11	39	19	167	96	144	37	36720

Table 3: Confusion matrices of one annotator against the other two for the data that have three annotators.

References

- Leon Derczynski, Diana Maynarda, Giuseppe Rizzo, Marieke van Erp, Genevieve, Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontchevaa. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.
- Tobias Ek, Camilla Kirkegaard, Håkan Jonsson, and Pierre Nugues. 2011. Named entity recognition for short text messages. *Procedia: Social and Behavioral Sciences*, 27:178–187.
- Sofia Gustafson-Capková and Britt Hartmann. 2006. Manual of the stockholm umeå corpus version 2.0: Description of the content of the suc 2.0 distribution, including the unfinished documentation by gunnel källgren.
- Janne Bondi Johannesen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdóttir, Anders Noklestad, Dimitris Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland scandinavian languages. *Literary and Linguistic Computing*, 20(1).
- Dimitrios Kokkinakis, Jyrki Niemi, Sam Hardwick, Krister Lindén, and Lars Borin. 2014. Hfst-swener – a new ner resource for swedish. In *Proceedings of the LREC workshop 'Beyond Named Entity Recognition: Semantic labelling for NLP tasks'*.
- Dimitrios Kokkinakis. 2004. Reducing the effect of name explosion. In *Proceedings of LREC*.
- J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.
- David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1).
- Ö. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.
- Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

From OCR to Digital Editions

Saranya Balasubramanian

Austrian Centre for Digital Humanities

Austrian Academy of Sciences, Austria

saranya.balasubramanian@oeaw.ac.at

Abstract

Documents digitised in mass-digitisation projects end up as high quality images and the text in them represented in one of the standard optical character recognition (OCR) formats. The Text Encoding Initiative (TEI) provides a much better way to encode the digitised content as it offers means to capture the metadata of the document and make detailed annotations. Since the OCR output only contains minimal markup that treats every isolated block of text as a paragraph, we developed models to automatically infer the structural markup and produce a richer TEI document. In particular we developed models to identify titles, subtitles, footnotes and page headers, and label OCR artefacts and surplus contents. In this paper we describe the capabilities of these models, our text encoding choices and the open challenges.

1 Introduction

The popularity of mass digitisation was at its peak in the 2000s which saw the Google books library project (Google, 2019) then known as Google Print, collaborate with several major national and university libraries to digitise their archives. Unfortunately much of the digital material simply remain in the information bank for search queries retrieving relevant sections and are not available for general perusal to the public due to copyright restrictions. Since then, optical character recognition (OCR) methods have come a long way, they are more accurate and scanning has become cheaper; and open sourced tools like Tesseract (Smith, 2007) have considerably improved, matching the quality of commercial tools like ABBYY Finereader. This has encouraged libraries to continue their digitisation efforts but it is information retrieval and not long-term electronic archival that remains the primary objective of these mass digitisation projects (Gooding, 2012).

In the project *Linked Cat+* we obtained 365,459 pages of digitised manuscripts published by the *Austrian Academy of Sciences* (ÖAW) from 1848 to 1917. These were also meticulously catalogued by *Bibliothek, Archiv und Sammlungen der ÖAW* (BAS:IS) giving high quality metadata for the contents of the digitised volumes. One of aims of this project was to archive these catalogued items into stand-alone digital documents – combining the text information from the relevant OCR pages and the metadata according to the guidelines provided by the Text Encoding Initiative or TEI (Giordano, 1994).

Combining multiple pages of OCR information into a single document at first seems trivial, requiring only to join the OCR contents and marking the beginning of every new page. However, producing a richer document with the content correctly marked up is a much deeper problem. The TEI export facility available in the Transkribus (Fronhöfer and Mühlbauer,) simply joins the page contents together; and in the case of dictionaries – a machine learning approach (?) is used to recognise the structure. Unlike dictionaries, the page structure of scholarly publications is much more variable, hence we developed a generalised solution to extract and mark up the content, which is demonstrated in the following sections.

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

2 Identifying structure in digitised texts

Formatted electronic texts usually embed markup information within them, which dictate the appearance of individual text elements (e.g. superscript, italics) and/or their role/position in the whole document (e.g. heading, paragraph). OCR outputs, specified in various XML-based formats, contain a limited amount of this information as well. It is limited in scope because the main objective of OCR engines is to recognise texts. They do recognise continuous blocks of text, but they do not differentiate their structural role in the document, marking them indiscriminately as paragraphs. Figure 1 show the “paragraphs” identified by the OCR engine Tesseract ¹.

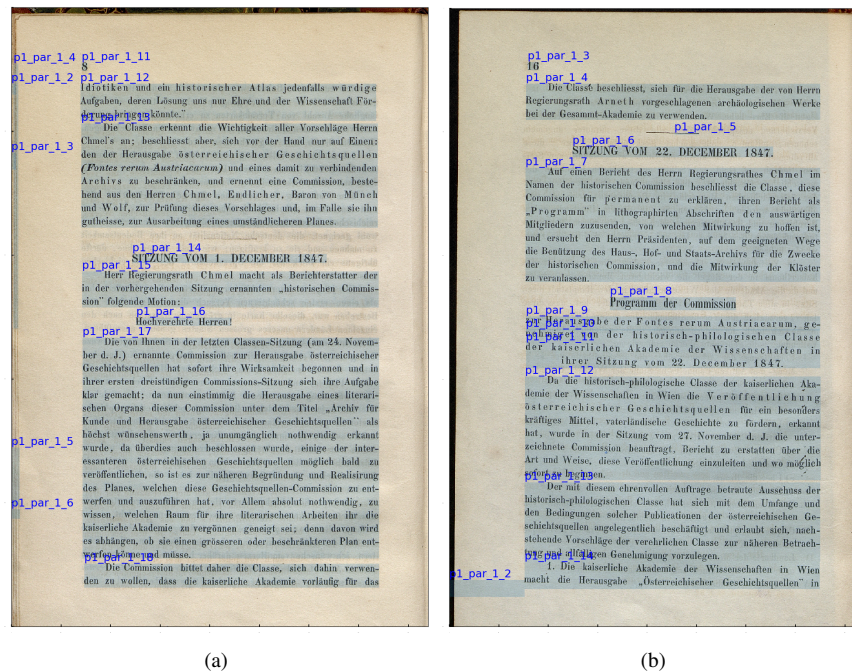


Figure 1: A sample of markup obtained from the OCR engine, highlighting the “paragraphs” found on two pages. Besides accurately segmenting the paragraphs on the pages, a few OCR artefacts near the left margins can be seen that have been misidentified as paragraphs.

The limitations of this approach are apparent from fig. 1; what appears to be blank left margin on both pages is wrongly indicated to contain paragraphs. These are artefacts of the OCR process. Structural elements on the page such as page numbers and headings are not technically “paragraphs” either but are labelled as such. Besides that, due to the page-centric nature of the OCR process, all text blocks are marked as ending on the pagebreak, even if they end mid-sentence and the actual paragraph continues on the following page. However, OCR engines provide the raw data of spatial information that nearly perfectly captures everything printed on a page; from the bounding box of every character/word, line and paragraph to the certainty with which it estimates its recognised text to be correct. Its accuracy, measured at word and character level is typically very high (Heliński et al., 2012); and this makes it an extremely reliable resource for post processing.

To deduce richer structural markup within a page and the continuation of content across pages, we have developed models that combine rule based and statistical approaches. The rules define expectations of the markup on a page – such as an offset at the first line of a paragraph, smaller size of font for footnotes and larger fonts or capital letters for headings, center aligned titles, etc. Statistical analysis is used to calculate the parameters used by these rules both locally (within a page) and globally (across a

¹tesseract 4.0.0-beta.1

set of pages) – such as the minimum and maximum size of the paragraph offset, etc. and identify outliers. For example, the OCR artefacts shown in fig. 1 can be filtered with a very high degree of accuracy by identifying those whose heights and widths and their positions lie outside the interquartile range for these values. These rules are not configurable at the moment but as the parameter values are statistically determined, the model is not tailored specifically to the pages in *Linked Cat+*. More methods are being tested and in development; presently the application offers the following capabilities in postprocessing OCR and generating a TEI document:

1. build a structured TEI document from multiple OCR pages,
2. add MARC21² data to the document header encoded as *teiHeader*,
3. identify and label OCR artefacts,
4. merge paragraphs that continue from one page to another,
5. recognise and mark up titles,
6. recognise and mark up page headers, footers and footnotes,
7. merge and split incorrectly formed paragraphs in OCR using offset analysis, and
8. identify beginning and end of the contents belonging to a document and mark the paragraphs found before the beginning or after the end as “surplus” (e.g. in fig. 1a: *p1_par_1_12* and *p1_par_1_13* are surplus).

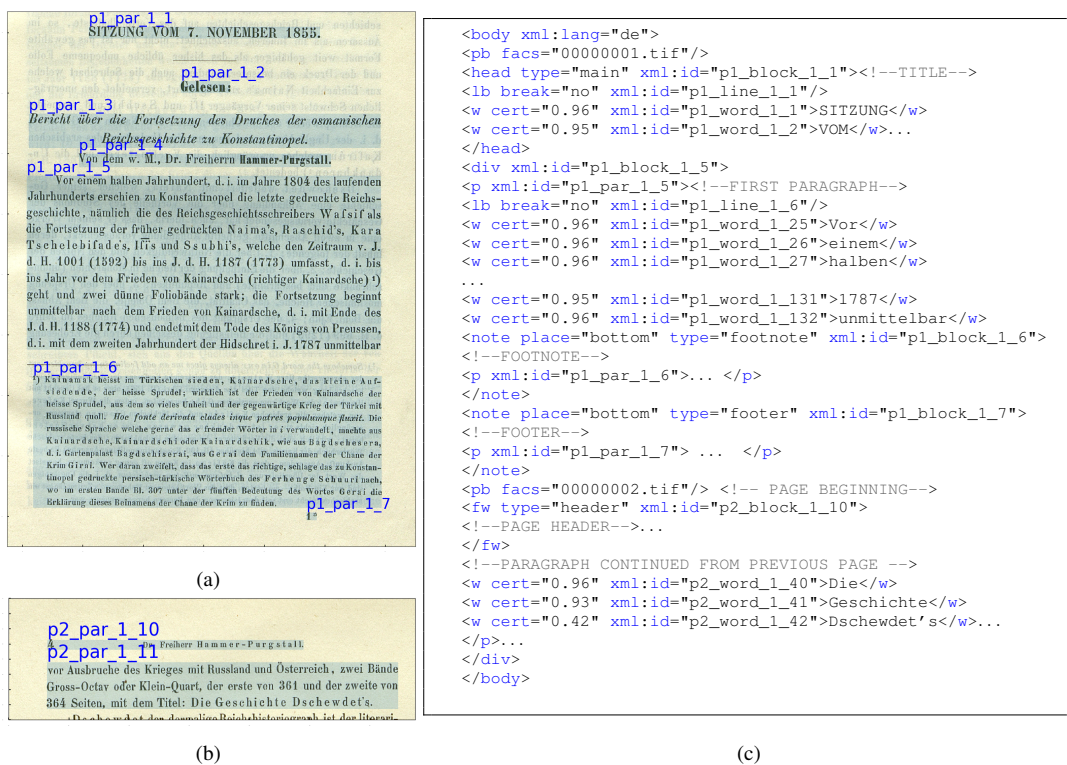


Figure 2: Pages demonstrating the continuation of a paragraph from one page to another, with footnotes and page headers “in between” (left) and the digitised text encoded in TEI format (right).

²MACHINE-Readable Cataloging – <https://www.loc.gov/marc/>

Figure 2 shows a sample of the encoded content – where the title, footnote, footer and page header are represented using the appropriate tags as recommended by the TEI without breaking the continuity of the paragraph in the body of the text; and the “first” paragraph on the second page is correctly identified as a continuation of the previous paragraph. Extensive use of footnotes is a particularly common occurrence in our catalogue and in scholarly publications in general. When the footnotes are treated as paragraphs in sequence as the OCR engines do, sequence of paragraphs in text loses its meaning. This is usually overcome by always presenting the digitised page coupled with the image of the page itself, from which a human reader can easily perceive markup and the flow of text. The aforementioned features reduce the coupling between the scanned pages and OCR and create an independent digital edition of the manuscript at least in straight-forward cases such as the examples show in figures 1 and 2. The next section discusses some more of the urgent problems that still exist.

3 Open challenges

Our approach with some basic assumptions about the page layout does have its limitations. Badly recognised texts cannot correct themselves in this transformation; in the future we would like to additionally provide an estimate of the quality of the digital edition we create, based on the quality of the source materials (in terms of the number of anomalies identified and fixed by the rules). Items in lists in print are usually recognised by OCR engines as separate paragraphs owing to the change in indentation and likewise, table cell regions are very well segmented but treated as paragraphs in the OCR formats. We are working to reliably identify and mark up these features.

Currently, the application has been tested only with the dataset of the *Linked Cat+* project but we plan to apply it to other datasets and generalize it. Furthermore, we aim to provide our application as a REST-based web service, freely usable by the community.

4 Conclusion

We reviewed some of the difficulties in digitising documents and presented automated methods for encoding scanned pages into a structurally coherent TEI document, considered an indispensable step towards creating good quality digital editions. We have showed potential uses for taking this additional step in digitisation projects and also highlighted the problems that still need to be solved. As we developed this primarily to analyse the historic scientific publications of a specific project and not extensively tested on other materials, we are cautiously presenting this work as “in-development” and we hope that together with the CLARIN developers’ community we can improve it further.

References

- A. Fronhöfer and E. Mühlbauer. Archivnutzung ohne Limit. Digitalisierung, Onlinestellung und das Projekt READ für barrierefreies Forschen. *Der Archivar, Zeitschrift fr Archivwesen*.
- Richard Giordano. 1994. Notes on Operations. The Documentation of Electronic Texts using Text Encoding Initiative Headers: An Introduction. *Library Resources & Technical Services*, 38(4):389–401.
- Paul Gooding. 2012. Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, 28(3):425–431.
- Google. 2019. Google Books Library Project. <http://www.google.co.uk/intl/en/googlebooks/library/partners.html>. Accessed: 2019-04-14.
- Marcin Heliński, Miłosz Kmieciak, and Tomasz Parkoła. 2012. Report on the comparison of Tesseract and ABBYY Finereader OCR engines.
- Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Enhancing Lexicography by Means of the Linked Data Paradigm: LexO for CLARIN

Andrea Bellandi, Fahad Khan, Monica Monachini
Institute For Computational Linguistics “A. Zampolli”
CNR Pisa, Italy
name.surname@ilc.cnr.it

Abstract

This paper presents a collaborative web editor for easily building and managing lexical and terminological resources based on the OntoLex-Lemon model. The tool allows information to be easily manually curated by humans. Our primary objective is to enable lexicographers, scholars and humanists, especially those who do not have technical skills and expertise in the Semantic Web and Linked Data technologies, to create lexical resources *ex novo* even if they are not familiar with the underlying technical details. This is fundamental for collecting reliable, fine-grained, and explicit information, thus allowing the adoption of new technological advances in the Semantic Web by the Digital Humanities.

1 Introduction and Motivation

Lexicography is traditionally recognised as that branch of applied linguistics which is concerned with the design and construction of resources that describe the lexicon of a language. In the digital era, it is very important that language resources can be represented in such a way that machines can process them and that they can be queried and shared across different communities. From this perspective it is possible to imagine a large-scale interconnected ecosystem of open, queryable and standardised lexicographic datasets and technologies. The Semantic Web, in particular the Resource Description Framework¹ (RDF), the Ontology Web Language² (OWL) and the Linked Data (LD) paradigm it is based on, makes this possible. Ontologies, in particular, have become an increasingly important method for formally modelling domains, and sharing them through the web.

In 2006, Tim Berners-Lee stated the four guiding principles for publishing data as LD³: i) use (Unique Resource Identifier) URIs as names for things; ii) use HTTP URIs so that people can look up those names; iii) when someone looks up a URI, provide useful information, using the standards (RDF, SPARQL⁴); iv) include links to other URIs, so that they can discover more things. These principles encourage both the maximum of interoperability between datasets and facilitate a more explicit encoding of meaning within and between datasets. The benefits of representing lexicographic content as LD are reusability, accessibility, interoperability and visibility at a Web scale. The content can be seamlessly integrated with content from external lexical resources. Lexical entries and their components are uniquely identified and become reusable thanks to URIs. The linguistic resource becomes a graph structure where each node is an entry point to navigate the whole graph, and each relation between two elements is typed and defined in a vocabulary.

In this context, the lemon model (McCrae et al., 2012), now called OntoLex-Lemon, was developed for creating lexicons that describe the lexicalization of ontological concepts. The number of users potentially

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

¹See <https://www.w3.org/RDF/> for RDF, and <https://www.w3.org/TR/rdf-schema/> for RDFS (last access: 19/04/2019).

²<https://www.w3.org/OWL/> (last access: 19/04/2019).

³If, in addition to the four prerequisites, data are made available under an open license, then it is classified as Linked Open Data (LOD).

⁴SPARQL query language for RDF <https://www.w3.org/TR/rdf-sparql-query/> (last access: 19/04/2019).

interested in editing or consuming OntoLex-Lemon data is thus very large (McCrae et al., 2017), and also the rapid development of the Linguistic Linked Open Data (LLOD) cloud⁵, is a success story influenced by the development of the OntoLex-Lemon model.

Our primary objective is to enable lexicographers, scholars and humanists to create lexical resources *ex novo* even if they are not familiar with the paradigm and the languages underlying their representation. Our tool allows information to be easily manually curated by humans. This is fundamental for collecting reliable, fine-grained, and explicit information. The tool we propose is among the first attempts to make the OntoLex-Lemon model accessible to all, especially to those who do not have technical skills in Semantic Web and Linked Data technologies, allowing the adoption of new technological advances in the Semantic Web by Digital Humanities.

2 Related works

Today, some tools for editing lexicons in different formats exist. Just to mention a few, we cite Lexus (Ringersma and Kemps-Snijders, 2007) and ColdicIn (Bel et al., 2008) for the Lexical Markup Framework (LMF) encoding, (Szymanski, 2009) for Wordnets, and CoBaLT (Kenter et al., 2012) for the management of lexicons in TEI P5.

In the context of the Semantic Web, editing tools for lexical or terminological resources are not so widespread, and in many cases, scholars are forced to adopt ontology editors, to formalize their lexical or terminological resources. Concerning the OntoLex-Lemon model, to the best of our knowledge, only two tools exist. The first one is lemon source, a Wiki-like site for manipulating and publishing lemon data aimed at the collaborative development of lexical resources. It makes it possible to upload a lexicon and share it with others. lemon source is an open source project, based on the lemon API, and it is freely available online for use. However, it deals with older versions of the OntoLex-Lemon model, and seems not to be updated anymore. The most relevant tool to ours is VocBench a web-based, multilingual, collaborative development platform for managing OWL ontologies, SKOS(/XL) thesauri, OntoLex-Lemon lexicons and generic RDF datasets. In (Fiorelli et al., 2017) the authors present their work on extending VocBench with facilities tailored to the OntoLex-Lemon model. However, LexO is more oriented on the needs of digital humanities. Firstly, we are working to link lexical senses to portions of text (e.g., attestations). Additionally, the editor is meant for formalizing peculiar features of linguistic resources such as etymology, representing aspects related to where words come from and how they originated, and diachrony for handling historical and ancient lexica and terminologies as well (Khan et al., 2016). It is worth emphasising that the process of extension in LexO is facilitated by the fact that OntoLex-Lemon, our lexical model of reference, is designed to be modular and to integrate new components easily.

3 LexO

Here, we present a first version of LexO⁶ (Bellandi et al., 2018), called LexO-lite⁷, that is a collaborative web editor for easily building and managing lexical and terminological resources for the Semantic Web and based on the OntoLex-Lemon model. The features of LexO were defined on the basis of our experience gained in the creation of lexical and terminological resources in the framework of several projects in the field of Digital Humanities. In particular:

- DiTMAO⁸, a digital-born multilingual medico-botanical terminology focused on Old Occitan and developed by philologists;
- FdS, a multilingual diachronic lexicon of Saussurean terminology in the framework of a lexicographic project⁹;

⁵<http://linguistic-lod.org/llod-cloud> (last access: 19/04/2019).

⁶The source code is available at <https://github.com/cnr-ilc/LexO-lite>. A simple demo of LexO adapted to a subset of Italian wordnet adjectives is available at <https://ilc4clarin.ilc.cnr.it/services/LexO>

⁷The suffix “lite” refers to the limited ability to manage small medium-sized lexica, due to the in-memory persistence we adopted, that is not a scalable strategy in case the resource size increases considerably. Currently, we are planning a “full” version of LexO for managing large resources.

⁸<https://www.uni-goettingen.de/en/487498.html> (last access: 19/04/2019)

⁹Demo available at <http://ditmao-dev.ilc.cnr.it:8082/saussure> (last access: 19/04/2019)

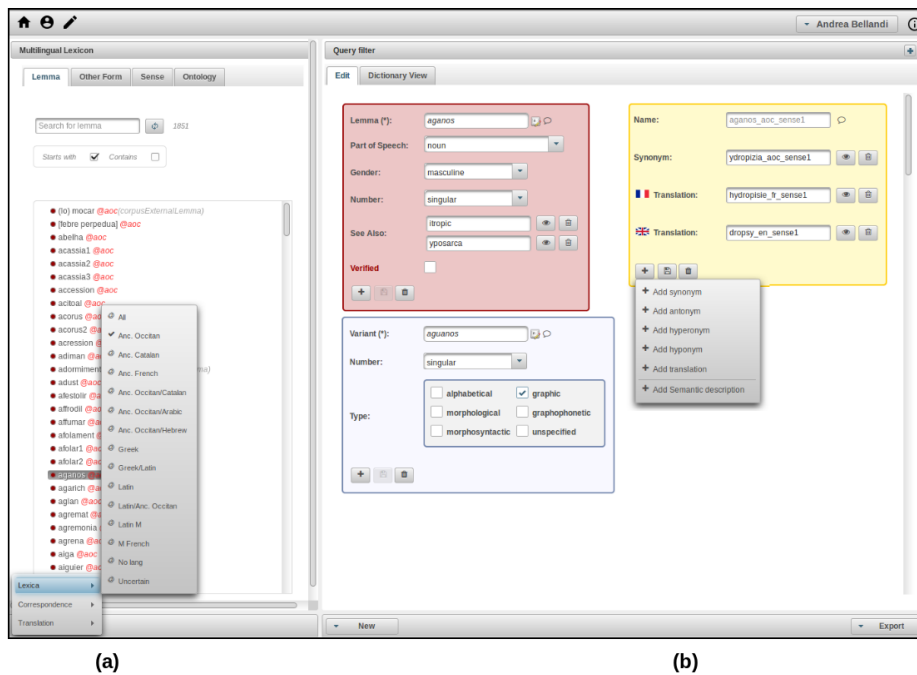


Figure 1: The main LexO's interface. (a) multilingual lexicon panel - (b) lexical entry editor.

- Totus Mundus, a bilingual Chinese-Italian resource dealing with Matteo Ricci's Atlas. LexO has been used by historians to build the linguistic resources related to the Map¹⁰.

During the development of LexO we have been influenced by some of the latest developments taking place in the European H2020 project ELEXIS (in which the Institute for Computational Linguistics of Pisa is involved as a partner). In particular our work has been closely informed by the survey of lexicographic users' needs conducted as part of the project and recently published as a deliverable¹¹. This is to ensure that the tool can be potentially used in as wide a range of lexicographic contexts as possible. Furthermore, LexO hides all the technical complexities related to markup languages, language formalities and other technology issues, facilitating access to the Semantic web technologies to non expert users. It provides possibility for a team of users to work on the same resource collaboratively each one according his/her own role(s) (lexicographers, domain experts, scholars, etc.). Finally, it provides a set of services implemented by means of the RESTful protocols that give software agents access to resources managed by means of LexO. The main interface of LexO, as shown in Figure 1, concerns the editing of a multilingual lexicon. It is mainly composed of 2 parts. The leftmost column allows scholars to browse lemmas, forms and senses, according to the OntoLex-Lemon core model, as Figure 1(a) shows. If the resource is multilingual, then users have the possibility of filtering lemmas, forms and senses by language. Information related to the selected entry is shown in the central panel where the system shows the lexical entry of reference, alongside the lemma (red box), its forms (blue boxes) and the relative lexical senses (yellow boxes), as shown in Figure 1(b). It is also possible to list the concepts belonging to an ontology of reference, and link lexical senses to them.

¹⁰Demo available at <http://lexo-dev.ilc.cnr.it:8080/TMLexicon> (last access: 19/04/2019)

¹¹See deliverable D1.1 "Lexicographic practices in Europe: a survey of user needs" at <https://elex.is/deliverables/> (last access: 19/04/2019).

4 The Potential Use of LexO in CLARIN and other Infrastructures

LexO fits in very well with a number of current CLARIN (as well as CLARIAH) initiatives and could prove itself a very useful tool within such an infrastructural context. For instance in the Netherlands in the context of CLARIAH, Linked Open Data has overtaken metadata publication in CMDI as one of the most important means of making data available¹². Recent Dutch initiatives have included the hosting of a workshop Linked Data for Linguistic Research¹³. In addition the Dutch Cornetto database¹⁴ and Open Dutch WordNet (ODWN)¹⁵ are both available as RDF. The Linked Open Data paradigm is also important for current CLARIN discussions respecting so-called Resource Families¹⁶ where it could become a core means of publishing lexicons. Linked Data has been also adopted in the ELEXIS e-lexicography infrastructure¹⁷ for facilitating linking and publishing of lexicons on the Web. Given the importance of Linked Data for lexicons then, it is clear that a tool like LexO could play a vital role in the editing and visualisation of such resources.

Acknowledgements

This work has been conducted in the context of the cooperation agreement between Guido Mensching, director of the DiTMAO project at the Seminar für Romanische Philologie of the Georg-August-Universität Göttingen, and the Istituto di Linguistica Computazionale “A. Zampolli” of the Italian National Research Council. The authors have also been supported by the EU H2020 programme under grant agreements 731015 (ELEXIS - European Lexical Infrastructure).

References

- Bel, N., Espeja Sergio, M. M. and Villegas, M. 2008. Coldic, a Lexicographic Platform for LMF Compliant Lexica. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).
- Bellandi, A., Giovannetti, E., and Weingart, A. 2018. Multilingual and Multiword Phenomena in a *lemon* Old Occitan Medico-Botanical Lexicon. *Information*, 9(3):52.
- Fiorelli, M., Lorenzetti, T., Paziienza, M. T., and Stellato, A. 2017. Assessing VocBench Custom Forms in Supporting Editing of Lemon Datasets In International Conference on Language, Data and Knowledge (pp. 237-252). Springer, Cham.
- Kenter, T., Erjavec, T., Dulmin, M. V., and Fier, D. 2012. Lexicon Construction and Corpus Annotation of Historical Language with the Cobalt Editor. In Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH '12, pp. 16.
- Khan, F., Bellandi, A., and Monachini, M. 2016. Tools and Instruments for Building and Querying Diachronic Computational Lexica. In Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH).
- McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. 2017. The OntoLex-Lemon Model: Development and Applications. In Proceedings of eLex 2017 conference, September (pp. 19-21).
- McCrae, J. P., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gmez-Prez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. 2012. Interchanging lexical resources on the Semantic Web. In Proceedings of Language Resources and Evaluation, 46(6), pp. 701709.
- Ringersma, J. and Kemps-Snijders, M. 2007. Creating Multimedia Dictionaries of Endangered Languages Using LEXUS. In Proceedings of Interspeech 2007. Baixas, France: ISCA-Int.Speech Communication Assoc., pp. 6568.
- Szymanski, J. 2009. *Wordventure developing wordnet in wikipedia-like style*.

¹²Jan Odijk, personal communication

¹³<https://www.clariah.nl/en/new/blogs/linked-data-for-linguistic-research#tuesday-7-february-2017>

¹⁴Cornetto: <https://portal.clarin.nl/node/1944>

¹⁵<http://wordpress.let.vupr.nl/odwn/>

¹⁶<https://www.clarin.eu/resource-families>

¹⁷<https://elex.is/>

Aggregating Resources in CLARIN: FAIR Corpora of Historical Newspapers in the German Text Archive

Matthias Boenig

Berlin-Brandenburg Academy of Sciences
and Humanities
Berlin, Germany
boenig@bbaw.de

Susanne Haaf

Berlin-Brandenburg Academy of Sciences
and Humanities
Berlin, Germany
haaf@bbaw.de

Abstract

Newspapers, though an important text type for the study of language, were not primarily part of the efforts to build a corpus for the New High German language carried out by the Deutsches Textarchiv (German Text Archive, DTA) project. After the finalization of the DTA core corpus, we started our efforts to gather a newspaper corpus for the DTA based on digital data from various sources. From the beginning, this work was done in the CLARIN-D context. Thanks to the willingness of external partners to pass on project results and to their cooperation it was possible to gather a corpus of historical newspapers, adapt it to a homogeneous set of guidelines and offer it to the community for free reuse. The poster is intended to provide insights into the newspaper and journal corpus of the DTA and to point out research possibilities which result from the aggregation of the digitized texts from various sources in the DTA.

1 Introduction

As a DFG long-term project, the German Text Archive (Deutsches Textarchiv; DTA, 2007–2019) pursued the goal of establishing the basis for a reference corpus for the development of the New High German language. However, for pragmatic reasons one type of text that was extremely important for the development, documentation, and dissemination of the German language had to be left aside for the time being: the newspaper. Since the completion of the work on the DTA core corpus, it is now possible to address this gap in the course of CLARIN-D and in connection with the “DTA Extensions” module (DTAE, 2019). Thanks to the willingness of external partners to pass on project results and their cooperation it was possible to gather a corpus of historical newspapers, adapt it to a homogeneous set of guidelines and offer it to the community for free reuse. In this poster, we want to present the corpora and discuss possible further enhancements and extensions.

2 Overview and General Workflow

The DTA newspaper corpus currently comprises 2,116 issues from more than 32 newspapers (cf. figure 1). It combines the diversity of the Mannheim Corpus of Historical Newspapers containing issues of several newspapers, with homogeneous corpora such as the “Neue Rheinische Zeitung” which concentrate on one newspaper and represent it as exhaustively as possible. The corpus is supplemented by journals such as the “Grenzboten” journal and the “Jahrbuch des Schweizer Alpen-Clubs” from the Text+Berg corpus.

With regard to their origin, the integrated newspapers and journals range from the 17th to the early 20th century, with the 19th and early 20th centuries as a core area. The corpus thus follows seamlessly on from the newspaper corpora of the DWDS, which in turn represent the 20th century in its course.

The respective newspaper and journal corpora originate from 6 different collaborations, with a total of more than 20 persons having been involved in the acquisition and preparation of the issues. This also shows the great potential that cooperation beyond project boundaries and the willingness to make research data available for other contexts bear.

The special value of aggregating these newspaper corpora into one corpus and integrating them into

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

the DTA is that the encoding and presentation of the respective newspapers is harmonized. This way, it becomes possible to explore and research all newspaper issues as a whole or separately without the need of further processing them or understanding the different ways of encoding.

The task of harmonizing newspapers from various sources and transferring them into one similar format is quite costly. Newspapers (esp. the newer ones) are extensively structured text types with many structural specialities to consider. For instance, articles of a respective rubric (e.g. political news) may be interjected by other articles of different rubrics (e.g. the feuilleton) or they may be continued in a later issue (Haaf and Schulz, 2014). At some point during our efforts to gather corpora of historical newspapers we defined a TEI tagset as a subset of the TEI format DTABf (Haaf et al., 2014; DTABf, 2019), to suit the special necessities of newspapers (Haaf and Schulz, 2014; DTABf-NP, 2019). This way, we were able to ensure homogeneous TEI tagging across the named newspaper corpora.

Once the texts were prepared according to the DTABf guidelines, the general DTA publication workflow was applied. First, the texts were integrated into the DTAQ platform where they underwent a quality assurance phase. In the course of corpus integration, the texts were automatically annotated with token-based linguistic information (lemma, POS, orthographically modernized form) in order to allow for elaborate corpus queries based on linguistic features (Jurish, 2012; Jurish et al., 2014). Subsequent to the quality assurance phase, the texts were integrated in the DTA platform where they are now freely available for exploration and download in different formats (TEI-DTABf, TCF, TEI with linguistic information, etc.).¹ Like all DTA texts the newspaper documents are also automatically integrated in the CLARIN infrastructure and can thus be found via VLO and explored via FCS.²

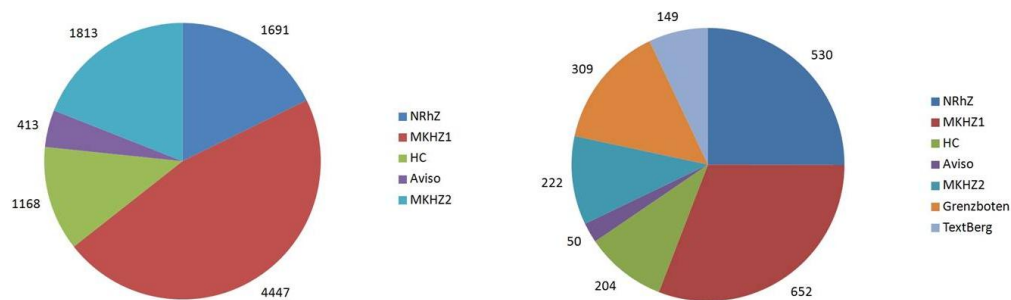


Figure 1: Newspaper Corpora within DTA by pages (left) and issues (right); pages are presented without Text+Berg (SAC, 87,006 pages) and Grenzboten (181,699 pages)

3 Sources

In the following, we present the newspaper and journal corpora gathered by the DTA and we offer insights on the workflows needed for their processing and harmonization.

3.1 „Aviso“

The oldest newspaper of the corpora presented here, is „Aviso. Relation oder Zeitung“ (Aviso, 1609), consisting of 51 issues dating back to the year 1609. It had been digitized in the course of a research project on newspaper language at the turn of the 16th to 17th century. For its integration into the DTA it was converted from the original transcriptions, carried out in a word processing software, into DTABf.

3.2 New Rhenish Newspaper

The „Neue Rheinische Zeitung“ (New Rhenish Newspaper; NRhZ, 1848/49) was provided by the long-term edition project „The Complete Works of Marx and Engels“ (MEGA, 2019). The newspaper appeared almost daily for one year during the German Bourgeois Revolution (1848/1849). The corpus

¹ See “Online Resources” section below for links. Cf. Geyken et al. (2019) for more information on corpus exploration and analysis options via DTA.

² Cf. Geyken et al. (2019) for more information on the DTA data preparation and publication workflow.

comprises the entire newspaper with all 301 issues plus supplements (531 documents). It had been transcribed manually using the TUSTEP format and converted semi-automatically to TEI according to the DTABf. All links between discontinuous parts of articles or sequels of articles were added manually (Haaf and Schulz, 2014).

3.3 Hamburg Correspondent

204 issues of the newspaper „Staats- und Gelehrte Zeitung des hamburgischen unpartheyischen Correspondenten“ and its predecessors (HC, 1712-1848) were manually transcribed and encoded according to the DTABf in the course of a cooperation project between the University of Paderborn and the BBAW. Since this newspaper was already digitized according to the DTABf it could be integrated into the DTA right away without the need of further harmonization steps.

3.4 Mannheim Corpus of Historical Newspapers and Journals

The first part of the „Mannheimer Korpus Historischer Zeitungen und Zeitschriften“ (Mannheim Corpus of historical newspapers and journals; MKHZ, 2019) was provided by the CLARIN centre at the Leibniz Institute for the German Language (IDS) in Mannheim. It comprises 652 issues of 21 newspapers which were originally transcribed using the TUSTEP format. The corpus was first converted into basic DTABf at IDS and then further processed and enriched with automatic steps complemented by manual efforts (e.g. adding information on linking among articles) at the CLARIN centre in Berlin. A second part of the MKHZ (236 issues of 6 newspapers) was then built up in the course of a cooperation project between BBAW and IDS (Fiechter et al., 2019). This way, digitization could immediately be carried out according to the DTABf, so that further steps of data conversion were not necessary.

3.5 „Grenzboten“

The journal “Die Grenzboten” (Grenzboten, 2019) was digitized at the SUB Bremen using OCR methods. As part of a cooperation project between BBAW and SUB Bremen, the BBAW carried out the manual pre-structuring and – based on the results – the automatic conversion of the OCR data into DTABf. In Bremen the automatic correction of the OCR data was carried out. The corpus comprises all 311 volumes of the journal.

3.6 „Text+Berg“

The Text+Berg corpus was built up by the equally named project at Zurich University (Text+Berg, 2008-2018). The data had originally been captured using OCR methods and was then manually corrected in the course of a community effort. Among other documents, it comprises multilingual journals of the Swiss Alpine Club. The 209 volumes were converted from their original TEI-XML format with additional stand-off annotations into the DTABf. The corpus was splitted in two parts which were included in the DTA corpus analysis platform “dstar” (SAC, 2019) and the DWDS (DWDS, 2019), according to their date of appearance and licencing constraints. Further enrichment and the integration of the respective volumes into the DTA platform are still in progress.

4 Conclusion and Further Work

The newspaper corpora presented above are all provided under free licenses and for free reuse by the research community. They have been or are being made available via the CLARIN infrastructure. A lot of effort went into the harmonization of the various formats the documents originally came in. They are now all encoded according to the TEI Guidelines and, more specifically, the DTABf, and are thus interoperable among one another.

Work on some of the corpora mentioned above is still in progress (i.e. MKHZ, part 1; Text+Berg). In addition, the conversion of one further corpus has started containing 367 issues of 27 newspapers, originating from a 17th century newspaper corpus.

In addition to these efforts, the collection is to be further enlarged as an opportunistic corpus in order to create the basis for a balanced newspaper corpus.

The poster is intended to provide an insight into the newspaper and journal corpus of the German Text Archive and to point out research possibilities which result especially from the consolidation of the digitized texts from very different sources in the DTA.

5 Acknowledgements

We want to thank all persons involved in the aggregation of the newspaper corpora presented here.

These are scholars and representatives of institutions willing to donate their research data to the DTA and CLARIN for further reuse by the community, namely (in alphabetical order): Prof. Dr. Noah Bubenhofer (University of Zurich), Prof. Dr. Stefan Engelberg (IDS Mannheim), Peter Fankhauser (IDS Mannheim), Prof. Dr. Thomas Gloning (University of Gießen), Dr. Jürgen Herres (BBAW Berlin), Prof. Dr. Michel Lefèvre (University of Montpellier), Manfred Nölte (SUB Bremen), Prof. Dr. Oliver Pfefferkorn (IDS Mannheim), Prof. Dr. Britt-Marie Schuster (University of Paderborn), Prof. Dr. Martin Volk (University of Zurich), Dr. Manuel Wille (University of Paderborn), Kay-Michael Würzner (SLUB Dresden).

Furthermore, we want to thank the (at the time) student collaborators at BBAW and cooperating institutions involved in the preparation of the newspapers for the DTA, namely (in alphabetical order): Maria Ermakova, Linda Feyerabend, Benjamin Fiechter, Rahel Hartz, Stefanie Hillmann, Amelie Meister, Nicole Postelt, Daniela Scharrer, and Stefanie Seim.

Reference

Articles

Benjamin Fiechter, Susanne Haaf, Amelie Meister, and Oliver Pfefferkorn. 2019. Presseschau um die Jahrhundertwende: Neue historische Zeitungen im DTA. In *Im Zentrum Sprache. Untersuchungen zur deutschen Sprache*. Blog article, February 6th. <https://sprache.hypotheses.org/1471>.

Alexander Geyken, Matthias Boenig, Susanne Haaf, Bryan Jurish, Christian Thomas, and Frank Wiegand. 2018. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In Henning Lobin, Roman Schneider, Andreas Witt, editors, *Digitale Infrastrukturen für die germanistische Forschung*, Berlin/Boston: pages 219–248.

Susanne Haaf, Alexander Geyken, and Frank Wiegand. 2014. The DTA “Base Format”: A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. *Journal of the Text Encoding Initiative*, 8.

Susanne Haaf and Matthias Schulz. 2014. Historical Newspapers & Journals for the DTA. In *LRT4HDA. Proceedings of the workshop held at the LREC’14*, May 26–31, 2014, Reykjavik (Iceland): pages 50–54.

Bryan Jurish. 2012. *Finite-state Canonicalization Techniques for Historical German*. PhD thesis, University of Potsdam. urn:nbn:de:kobv:517-opus-55789.

Bryan Jurish, Christian Thomas, and Frank Wiegand. 2014. Querying the Deutsches Textarchiv. In Udo Kruschwitz, Frank Hopfgartner, Cathal Gurrin, editors, *Proceedings of the Workshop “Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities” (MindTheGap-2014)*, March 4, 2014, Berlin (Germany): pages 25–30.

Online Resources

Aviso. 1609. *Relation oder Zeitung*. Wolfenbüttel, 1609. In Deutsches Textarchiv http://www.deutschestextarchiv.de/anonym_aviso_1609.

DTA. 2007–2019. *Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*. <http://www.deutschestextarchiv.de/>.

DTABf. 2019. *DTA Base Format*. <http://www.deutschestextarchiv.de/doku/basisformat/>.

DTABf-NP. 2019. *DTABf for Newspapers*. <http://www.deutschestextarchiv.de/doku/basisformat/zeitung.html>.

DTAE. 2019. *Deutsches Textarchiv – Extensions*. <http://www.deutschestextarchiv.de/dtae/>.

DWDS. 2019. *Digital Dictionary of the German Language*. <https://www.dwds.de/>

Grenzboten. 2019. *Die Grenzboten*. <http://deutschestextarchiv.de/doku/textquellen#grenzboten>.

HC. 1712–1848. *Hamburgischer Correspondent*. <http://deutschestextarchiv.de/doku/textquellen#correspondent>.

MEGA. 2019. *The Complete Works of Marx and Engels*. <http://www.bbaw.de/en/research/mega>.

MKHZ. 2019. *Mannheim Corpus of Historical Newspapers and Journals*.

<http://deustextarchiv.de/doku/textquellen#mkhz>.

NRhZ. 1848/1849. *Neue Rheinische Zeitung*. <http://www.deustextarchiv.de/nrhz/>.

SAC. 2019. *Journal of the Swiss Alpine Club, part of Text+Berg corpus*. <http://kaskade.dwds.de/dstar/textberg/>.

Text+Berg. 2008–2018. *Text+Berg digital*. <http://textberg.ch>.

CLARIN and Digital Humanities. A Successful Integration

Elisabeth Burr

Institut für Romanistik
Leipzig University
Germany

elisabeth.burr@uni-
leipzig.de

Marie Annisius

Institut für Informatik
Leipzig University
Germany

marie.annisius@uni-
leipzig.de

Ulrike Fußbahn

Institut für Romanistik
Leipzig University
Germany

ulrike.fussbahn@uni-
leipzig.de

Abstract

The collaboration between the European Summer University in Digital Humanities “Culture & Technology” (ESU) and CLARIN-D is a concrete example of the successful integration of a digital language resources and technology research infrastructure for the humanities and social sciences and Digital Humanities. While a thorough analysis of this collaboration, its outcome and its impact need to be postponed to a later date, we would like to offer at least some insight into this collaboration. In the first part of our presentation, we will outline briefly the foundation and specific nature of the ESU. The second part will explain how the collaboration between the ESU and CLARIN-D came about and what it consists of. The third part is dedicated to results and tries to draw a few conclusions before it expresses some hopes for the future. As the proposal was not discussed with colleagues from CLARIN-D, the view on this collaboration is a personal and partial one.

1 Introduction

Training initiatives in Digital Humanities have been around for quite some time. Furthermore, over the last two decades many Summer, Winter, Spring and Autumn schools in Digital Humanities have taken place. Most of them remained, however, one-time events.¹

The oldest initiative is the *Digital Humanities Summer Institute* (DHDI) at the University of Victoria in Canada, which was founded in 2001 and has managed to raise the number of its participants from 35 to over 900. The other long-lasting Digital Humanities Summer Schools were founded towards the end of the last century: The *Digital Humanities at Oxford Summer School* (DHOxSS) started in 2008 as a TEI Summer School and became a Digital Humanities Summer School in 2011. The *IDE Summer School* of the Institut für Dokumentologie und Editorik took place for the first time in 2008. The *European Summer University in Digital Humanities “Culture & Technology”* (ESU) and the *Edirom Summer School* (Edirom) were founded in 2009.

While some of these Summer Schools are devoted to specific Digital Humanities,² Summer Schools like DHDI, DHOxSS and the ESU cater for the diversity of the field. The only Summer School which explicitly aims at integrating language resources, Linguistics and Digital Humanities is, however, the Leipzig Summer University.

In the following I will outline briefly the foundation and specific nature of the ESU before I explain how the collaboration between the ESU and CLARIN-D and lately also the CLARIN-ERIC came about

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

¹ To my knowledge a list of all the DH Summer, Winter, Spring and Autumn Schools does not exist, the IfDH|b has, however, put together quite an impressive collection of such initiatives which spans the years 2012 (incomplete) to 2018.

² See for example the *IDE Summer School* and the *Edirom Summer School*, which focus on digital editions in general and music editions in particular.

and what it consists of. In the last section we will not only hint at some results and draw preliminary conclusions but also express some hopes for the future.³

2 The European Summer University in Digital Humanities (ESU)

In the winter semester 2007/2008 King's College London (UK), Debreceni Egyetem (Hungary), Universität Leipzig (Germany) and Oulun Yliopisto (Finland) taught a *Joint seminar programme Culture & Technology* together via video conference. In the winter semester 2008/2009 the experiment continued with 8 European Universities: Universidad de Alicante (Spain), Università Bologna (Italy), Debreceni Egyetem (Hungary), University College Dublin (Ireland), University of Glasgow (Scotland), Universität Leipzig (Germany), King's College London (UK) and Oulun Yliopisto (Finland). Furthermore, in December 2007 a workshop "Text Markup & Database Design", sponsored by the *Association for Literary and Linguistic Computing (ALLC)*⁴ took place at the University of Leipzig. The results were very encouraging.

2.1 The foundation years of the ESU

In 2009 the *European Summer University in Digital Humanities (ESU)* was founded, which as "Culture & Technology" indicates, saw itself as a follow up to the *Joint European seminar programme*. The stance taken towards the working languages was also European. While English for pragmatic reasons had to be the lingua franca, tendencies towards a more and more monolingual culture of science were to be countered by showing concrete esteem and respect for European multilingualism and the variety of European knowledge cultures. Thus, other languages were also to be used during the workshops as long as everybody was included.⁵

Three times the ESU was able to rely on the support of the *Volkswagen Foundation* (2009, 2010 and 2012). The number of participants went up from 31 to 60, and the workshops from 4 to 6. Via the participants of these three editions, 27 different countries were represented. Already during the application phase for the 2nd ESU it became clear that the Summer University was seen as an institution in Europe, and not as an institution promoting and developing Digital Humanities within Europe as we had seen it originally.

2.2 Specificity of the ESU

The ESU is different from other training initiatives in that it does not aim at growing and attracting large numbers of participants. Instead, it aims at keeping the number of participants at around 60. It also does not look at the participants as paying entities or as a means to make profit. Instead, it tries to see every participant as an individual with a particular background and specific needs and as a potential member of an international community which fosters the building of networks, collaboration, exchange and understanding across borders of all kinds (disciplinary, linguistic, cultural, national, status and the like). Therefore, people interested in participating cannot simply register, but need to apply for a place, handing in a CV and a motivation letter in which they specify the workshop(s) of their choice. All applications are read by the director, who then assigns them to the experts offering the respective workshop, and to members of the Scientific Committee for review. The director thus has a very good insight into the composition of the group of applicants, which changes every year. The director also functions as a gatekeeper as, instead of assigning all applications directly to the reviewers, s/he tries to make sure that an application is serious and that the Call has not been (willingly) misunderstood. Reviewers are thus not unnecessarily bothered with problematic applications, the number of which can be quite high.

The aim of the reviewing process is not to divide the applications into 'good' and 'bad' ones. Instead, reviewers are supposed to act as mentors. As a result of their expertise and experience, they can provide valuable feedback on the choice of workshop(s), on the soundness of expectations and on the well-foundedness of research questions. Such feedback is taken very seriously by the applicants.

³ We will need to name some individuals, because without them this collaboration would not have come about. The many others who filled this collaboration with life, helped to overcome difficulties and fostered its continuous growth, will, however, have to remain anonymous, as a thorough analysis of this collaboration is not possible in this place.

⁴ The ALLC is now *The European Association for Digital Humanities (EADH)*.

⁵ See <http://esu.culintec.de/?q=working_languages>.

Obviously, the reviewing of applications means a lot of work for the experts long before the start of the ESU, but it also gives them the chance to get to know the participants of their workshop beforehand and to take their needs and expectations into account when they prepare it. It also means that nobody arrives at the ESU as an anonymous entity.

3 ESU and CLARIN

2013 was a particularly difficult and memorable year. The DAAD had not accepted our proposal for funding and the ESU was in danger of having to be cancelled. Some small financial support arrived, however, from the Friends of Leipzig University, DARIAH-DE sponsored a lecture, and the International Office of Leipzig University opened its East-European Partner Universities bursary scheme to participants of the ESU. At some point it became clear, that the biggest and decisive support would come from CLARIN-D and that the ESU was safe. In fact, CLARIN-D was not only willing to sponsor a workshop,⁶ but also to grant funding, which would allow us to hand out tuition fellowships to 17 of the altogether 39 participants of the 4th ESU. This support was made available by the “Training and Education” WP, chaired in those days by Elke Teich.

3.1 The first joint Summer University

If it had not been for Elke and her team, the collaboration could have ended here. At the beginning it seemed, in fact, that any further collaboration would mean that the ESU became a CLARIN-D Summer School. Three facts helped to overcome this difficult situation. The meetings Elke organised on the 10th October 2013 in Mannheim at the IDS together with quite a number of CLARIN-D representatives, the 13th of January 2014 in Leipzig, and the 27th of June 2014 again in Mannheim brought the two somehow akin but also very different communities together. The 2 years funding granted the ESU in January 2014 by the DAAD made it possible, furthermore, to construct a partnership where both parties could work together at eye level. Last but not least, the wiki the Saarbrücken team set up allowed for a transparent and cooperative planning of the 2014 Summer University.⁷ Due to the active involvement and support of Elke and Hannah, her assistant, also the radical change from an ESU, which up to then had offered exclusively workshops which ran through all the nine days, to an ESU, which would accommodate also workshops which lasted only 5 days, could be managed. All the five workshops CLARIN-D wanted to offer could, in fact, take place and the *Joint European Summer University Digital Humanities & Language Resources* with its 61 participants from 24 countries of the world became a real success.

3.2 Workshops contributed by CLARIN⁸

Since 2014 the joint ESU runs through 11 whole days. At every ESU of the last 6 years 11 workshops could be realised. Between 3 and 5 of them were offered by colleagues from the following CLARIN-D centres: Universität Saarbrücken, IDS Mannheim, Berlin-Brandenburgische Akademie der Wissenschaften, Universität Tübingen, LMU München, Universität Hamburg, Universität Leipzig, and Universität Stuttgart.

By means of these workshops the linguistic, language resources or specific CLARIN perspective could be integrated with the more traditional Digital Humanities perspectives. Quite naturally, written, spoken, and historical corpora, their creation (compilation, digitization, annotation, quality assurance) and analysis (querying, comparing, pattern searching) and the respective tools always played a very prominent role. Another component, which CLARIN workshops contributed to the Summer University was data modelling and the usage and querying of (graph) databases as well as data management and its legal and ethical issues. Obviously CLARIN services were also introduced to the community. CLARIN workshops on spatial analysis had such a strong impact on the ESU community that a new way of thinking about humanities data, their curation and their use in digital mapping environments evolved. Topics

⁶ According to the mails I went through, Andreas Witt, who knew the ESU from personal experience having taught a workshop already in 2012, and Gerhard Heyer were the driving forces behind the sponsoring of the first CLARIN-D workshop ever offered at the ESU.

⁷ The *Joint ESU DH C & T and CLARIN-D Summer School* wiki is still available.

⁸ As funding for the CLARIN-D involvement with the ESU at some point started to be granted by the CLARIN-ERIC I will in the following just refer to CLARIN.

like programming for the Web, rule-based machine translation and neural networks for NLP were also contributed by CLARIN.

3.3 Fellowships

Starting with the 6th ESU, which happened in 2015, and until the 9th ESU, which took place in 2018, CLARIN-D, on top of the workshops, offered also around 10 tuition fellowships. To apply for such a fellowship, a written application (ca. 500 words) in English or German, including a short statement on the research interest, on the status, the university affiliation and the reason for applying for such a fellowship had to be submitted via ConfTool. The selection of the fellows was done by a committee consisting of CLARIN-D representatives, and CLARIN representatives also handed out the fellowship certificates at the opening of the ESU.

3.4 Some Results

From 2009 to 2019 851 people, including 269 workshop leaders and lecturers, from 60 different countries of the whole world participated in the *European Summer University in Digital Humanities “Culture & Technology”*. The great majority of them was part of the joint endeavour of CLARIN-D and the ESU.

For the colleagues from the CLARIN-D centres this meant being confronted with a very international group of people who did not always speak English perfectly, who brought with them many different languages and very different cultural backgrounds, had very diverse needs and expectations and could not always be contented with German corpora or tools, or were sometimes very critical with respect to the lack of attention granted to the TEI. They thus really had the chance to meet the user community for whom CLARIN intends to work.

Many of the participants who were interested in taking part in CLARIN workshops arrived with a linguistic research question in mind. Their aim at first was only to acquire relevant methods and skills or get to know useful tools, which could help to solve this question. Their interest in the Digital Humanities was limited. When they left the Summer University, however, they had acquired a much broader perspective and could see that all these methods, skills and tools can play not only a role in purely linguistic projects or research but can also be integrated with literary, cultural (heritage), historic and other research questions. Furthermore, they had the chance to form networks across the borders of the CLARIN and Digital Humanities communities, across disciplines, countries, languages and cultures and had been confronted with overarching methodological and theoretical questions. The ESU is after all not only about skills, methods, and tools, but about the Digital Humanities. The fact that all that was learnt during the individual workshops could be seen in this much broader perspective, was guaranteed by the various lectures and project presentations and the panel, which are an integral part of ESU's programme.

4 Conclusion

As can be seen in the videos about the ESU 2015 in YouTube, in the article about the ESU 2017 in the CLARIN-D blog, and the reports that winners of CLARIN fellowships wrote about the experiences gathered at the ESU 2018, as well as the many positive remarks about former ESUs we find in applications for a place at the Summer University, the integration of the ESU and CLARIN is not only a real success for the German and *European Research Infrastructure for Language Resources and Technology* but also for Digital Humanities, because it helps to pull down the borders between the more language(s) and the more literary and culture oriented sides and contributes to DH projects becoming even more open to very diverse research questions. It also brings the infrastructure into very direct contact with the users, allows for intensive exchange and encourages extensive learning processes. It is to be hoped that the steps already taken towards the institutionalisation of the ESU at the University of Leipzig will be granted success and that the fruitful collaboration between CLARIN and the ESU will continue for more years. As one Digital Humanities expert put it lately, if the ESU does not go on then there will be no serious Digital Humanities training event outside of North America anymore and Europe will lose the strong impact on the formation of digital humanists which has been built up over the years.

References

- CLARIN-D (2015): *ESU European Summer School for Digital Humanities, Leipzig 2015* <https://www.youtube.com/watch?v=Oh7UJi_2118> [26.08.2019].
- CLARIN-D (2015): *ESU Sommerschule für digitale Geisteswissenschaften, Leipzig 2015* <<https://www.youtube.com/watch?v=yDfeTMHt9Q4&feature=youtu.be>> [26.08.2019].
- CLARIN-D (19.10.2017): „Digital Humanities studieren und netzwerken mit CLARIN“, in: *CLARIN-D Blog* <<https://www.clarin-d.net/de/blog-clarin-d/7-digital-humanities-studieren-und-netzwerken-mit-clarin>> [29.04.2019].
- CLARIN-D (n. y.): „Erfahrungsberichte zur European Summer University (ESU) 2018 online“, in: *CLARIN-D* <<https://www.clarin-d.net/de/aktuelles/475-esu-erfahrungsberichte>> [26.08.2019].
- European Summer University in Digital Humanities “Culture & Technology” <<http://esu.culintec.de/>> [26.08.2019].
- If|DH|b = Interdisziplinärer Forschungsverbund Digital Humanities in Berlin (n. y.): „Überregionale DH Summer Schools“, in: *If|DH|b* <<http://www.ifdhberlin.de/lehre/dh-summer-schools-ueberregional/>> [26.08.2019].
- Wiki: *Joint ESU DH C & T and CLARIN-D Summer School* (10.06.2014) <<http://www.clarind.uni-saarland.de/dokuwiki/doku.php?id=start>> [26.08.2019].

AcTo: How to Build a Network of Integrated Projects for Medieval Occitan

Gilda Caïti-Russo, Hervé Lieutard

Laboratoire LLACS
Univ Paul-Valéry Montpellier 3
gilda.russo@univ-montp3.fr
herve.lieutard@univ-montp3.fr

Jean-Baptiste Camps

Centre Jean-Mabillon
École nationale des chartes
Université PSL, Paris
jean-baptiste.camps@chartes.psl.eu

Gilles Couffignal

Université Paris-Sorbonne
gilles.couffignal@paris-sorbonne.fr

Francesca Frontini

Laboratoire PRAXILING
Univ Paul-Valéry Montpellier 3
francesca.frontini@univ-montp3.fr

Elisabeth Reichle

Ludwig Maximilian University of Munich
Elisabeth.Reichle@dom.badw.de

Maria Selig

Universität Regensburg
Maria.Selig@sprachlit.uni-regensburg.de

Abstract

We present AcTo, a network of integrated projects for the development of language resources and tools for Medieval Occitan. This abstract illustrates the resources in the network, as well as the first steps towards their integration, aiming towards the harmonisation and interoperability of NLP and lexical resources for the annotation of digital editions.

1 Introduction

Computational linguistics methods and digital language resources are becoming more and more important for philology. Computational philology approaches and infrastructures develop and adapt tools and methods specifically for the needs of scholars working with historical languages, and the development of computational corpora and lexicons is flourishing in this domain (Crane, 2012; Passarotti et al., 2019). Medieval philologists are not lagging behind digital classicists in the development of new approaches and solutions, with successful experiments in the application of OCR and textual analysis techniques to their manuscripts (Pinche et al., 2019). We concentrate here on computational philology approaches for Medieval or Old Occitan. While being the ancestor of a modern language spoken by minorities in France, Spain (Catalonia), and Italy, Medieval Occitan is also, and crucially, the language of a corpus of texts fundamental for the pre-modern cultural history of Europe. Indeed the corpus of Old Occitan literature, and especially the texts of the Troubadours have had a great influence in the development of modern European literature and beyond¹. As proof of this, Medieval Occitan is taught and studied at academic level in many European and American universities, and beyond.

In this abstract we illustrate the first activities of AcTo², a network of data and resource centers for the study of Medieval Occitan headed by Université Paul-Valéry in Montpellier, France, which gathers together projects from different countries (France, Italy, Spain, Germany, UK)³. The aim of the project is to federate existing resources (digital editions, lexicons, but also tools), harmonising the data and metadata

¹This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

²It would be impossible to trace here the influence of Occitanism in literature. Troubadours have been dubbed the inventors of modern verse (Wilhelm, 1970) and their influence has extended to contemporary American authors such as Ezra Pound and more recently W.S. Merwin, poet winner of the Pulitzer price for poetry in 2008.

³AcTo stands for *Acolhir e Tornar*, “to collect and return”, a line from troubadour Guiraud de Bornèlh.

⁴See the project’s web site for the complete list, as well as descriptions of individual projects and proceedings of past meetings (Caïti-Russo et al., 2019).

encoding within the projects as well as with international standards⁴. During a first project meeting, several working groups were constituted, dedicated to the alignment of metadata and documentation⁵, to the annotation and referencing of place and persons' names in digital editions, and to legal issues. The project plans to draw experience and help from CLARIN, Huma-Num and other infrastructures in order to align itself to existing best practices.

Within this framework, one specific effort, which constitutes the main object of this abstract, concerns orthographic normalisation across projects. This is a particularly important issue, since Medieval Occitan orthography was not standardised. Digital editions, while preserving the verbatim transcription, should also allow for search by normalised forms. Harmonising the normalisation as well as the lemmatisation choices is a crucial pre-requisite for a federated search throughout all existing corpora. Here we shall illustrate ongoing activities which have the aim of automatically linking a lexicographic resource to an existing corpus by means of an ad hoc morphological analyser, which automatically reconverts non standard forms to the normalised lemma.

2 The resources

2.1 The Thalamus project and corpus

The Thalamus ANR project carried out the digitising and TEI encoding of the manuscript corpus of the government books of the medieval city of Montpellier. The critical digital edition is available online (Carrasco et al., 2014...), with the various manuscripts displayed in parallel, aligned by year, something which allows scholars to investigate how successive chronicles have re-written and edited past events of the city in the light of contemporary matters. This synoptic edition makes it possible for scholars to study the diachronic evolution of pre-diglossic Occitan from 1260 to 1426 as no other document can do. So far the normalisation and annotation of the text has been limited to place and person names, which are searchable from two dedicated indices, independently from their written form, which may vary. The current objective is to implement a search by forms and lemmas, in order to manage and study graphical variation. For this reason we are currently looking into making the TEI Thalamus corpus, the Medieval Occitan dictionary (*Dictionnaire de l'occitan médiéval*, see 2.2) and the OMÉLiE project (2.3) all interoperable.

2.2 The DOM, a reference lexicon for Medieval Occitan

The *Dictionnaire de l'occitan médiéval* (DOM), is a project coordinated by the Bayerische Akademie der Wissenschaften. It is a reference lexicographic resource for Medieval Occitan philologists. Based on PostgreSQL, the DOM is available online (Stempel et al., 1996...). The lexical entries, completed with bibliographic references, list the lemmas and all of their variants, the (polysemic) meanings, and a list of attestations. The dictionary provides a separated alphabetic list of lemmas and variants, so that search is also possible by all of the variants. The articles are connected by hyperlinks to the *Französisches Etymologisches Wörterbuch* (FEW) for further etymological research (Wartburg, 1922 1967). Linking editions to the DOM has been identified as an important task in the overall goal of federating the various digital editions projects within AcTo. The DOM will provide the necessary lexicographic information for normalisation and lemmatisation of Old Occitan texts when digitalised and prepared for annotation. DOM entries are provided with a unique URI, which could be used as a unique lemma reference by digital editions of texts. The DOM lexicon therefore offers the opportunity of creating a platform for access, analysis and interpretation of Old Occitan digitised texts. It facilitates cross-collection search inside the Old Occitan corpus and offers the possibility to make use of the existing network of lexicographic sources provided by the DOM. In the long run, the aim is to create a hybrid system of lexicographic devices and electronic corpora ("*Digitales lexikalisches System*") hosted by the Bayerische Akademie der Wissenschaften/Leibniz Rechenzentrum.

⁴The project focuses on the Medieval stage of the Occitan language, but in a diachronic perspective the relationship to modern Occitan is crucial; AcTo is supported by AIEO (Association Internationale d'études Occitanes) and by the CIRDÒC (Centre Interrégional de développement de l'occitan), which maintains a repository, *Occitanica.eu*, hosting a number of language resources for the Occitan language and will be eventually related to *Lo Congrès*, the most important repository of language resources for Modern and Contemporary Occitan (<https://www.locongres.org>).

⁵Discussions over these aspects will be headed by the CIRDÒC, a partner in AcTo; notice that the *Occitanica.eu* datacenter is already harvested by Europeana; harvesting to the VLO of relevant resources is currently being considered.

Thalamus.test

Quick links

- Search tokens
- Correct tokens
- Last corrected tokens
- Export tokens
- Corrections history
- Control List
- Editions history

Correct tokens with

- Unallowed lemma
- Unallowed POS
- Unallowed morph

Corpus Thalamus.test - List of tokens

1 2 3 4 5 ... 22 23

Id	Form	Lemma	POS	Morph	Context	Similar	Save	+
1	L'	lo2	DETdef	NOMB.=s GENRE=m CAS=r	L' an MLXXXVIII , los crestians prezeron Barsalona . L'	2	Save	+
2	an	an	NOMcom	NOMB.=s GENRE=m CAS=r	L' an MLXXXVIII , los crestians prezeron Barsalona . L' an	15	Save	+
3	MLXXXVIII	@num@	ADJcar	NOMB.=s GENRE=m CAS=r	L' an MLXXXVIII , los crestians prezeron Barsalona . L' an MCLXXXII	0	Save	+
4	,	,	PONtbl	MORPH=empty	L' an MLXXXVIII , los crestians prezeron Barsalona . L' an MCLXXXII ,	109	Save	+
5	lcs	lo2	DETdef	NOMB.=p GENRE=m CAS=n	L' an MLXXXVIII , los crestians prezeron Barsalona . L' an MCLXXXII , a	16	Save	+
6	crestians	crestian	NOMcom	NOMB.=p GENRE=m CAS=n	L' an MLXXXVIII , los crestians prezeron Barsalona . L' an MCLXXXII , a XIII	0	Save	+
7	prezeron	prèndre	VERcig	MODE=ind TEMPS=ppp PERS.=3 NOMB.=p	L' an MLXXXVIII , los crestians prezeron Barsalona . L' an MCLXXXII , a XIII setembre	0	Save	+

Figure 1: The Pyrrha post-correction interface.

2.3 The OMÉLiE project

OMÉLiE (*Outils et méthodes pour l'édition linguistique enrichie*) is a project of the *École des chartes* in Paris, with support from SCRIPTA (Université PSL) and the DIM *Sciences du texte et connaissances nouvelles* (Région Île-de-France). Its objectives are to offer tools and methods for the linguistic enrichment and analysis of ancient and medieval texts. Currently the research concentrates on Old French and Occitan. The aim is to offer an environment in which TEI editions can be uploaded, automatically lemmatised and annotated with morphosyntactic tags (Part-of-Speech, morphological analysis) and later post-corrected by humans to produce better models. The annotation system is based on deep learning methods, and in particular uses the *Pie* tagger (Manjavacas et al., 2019). It is integrated in a post-correction environment, Pyrrha (Clérice et al., 2019), that allows for close inspection as well as batch corrections, and can handle reference lists of lemmas and tags (see Figure 1). Both are available as open source software. This environment, which had initially been tested on Medieval French corpora, has now been applied to Medieval Occitan texts such as the romance of *Flamenca* and, crucially, the *Thalamus*. Pyrrha could easily be extended to support more recent varieties of Occitan, by ensuring interoperability with the CORLIG project (*Corpus de la Renaissance Littéraire gasconne*) coordinated by the Sorbonne University in Paris.

3 The lemmatisation project

The current project aims at bringing together DOM and OMÉLiE for an improved lemmatisation of the *Thalamus*. A first lemmatisation strategy has been developed, which is based on *LemmaGen* (Juršić et al., 2010) learning lemmatisation rules from existing lemma-wordform pair examples extracted from the DOM articles. In order to improve on that, and to provide for the full morphological analysis and lemmatisation of word forms, an annotation campaign is currently being carried out by the *Thalamus* and the OMÉLiE teams, to create an annotated corpus using Pyrrha. The annotation is performed by correcting the output of a first basic model, and will serve as training and test set for the creation of a better one. Following the annotation guidelines, first the lemmatisation is corrected, strictly following the DOM orthography; missing lemmas are recorded and set aside for their integration in the DOM; then the morphosyntactic annotation of the token in context is carried out, using the *Cattex* tagset (Prévost et al., 2013; Guillot et al., 2013). The poster presentation will show the first results of the lemmatisation model, and show how the link between the digital edition and the DOM can be encoded in the TEI edition.

4 Future work

Due to its influence that goes well beyond the borders of historical Occitania and modern day France, Medieval Occitan can be seen as part of a shared European heritage. For this reason we intend to integrate the AcTo community within the activities of CLARIN ERIC, as well as those of various national consortia, in order to ensure the visibility and interoperability of our digital resources as well as to exploit and adapt existing solutions and technologies. Future objectives of AcTo are:

- to make TEI editions of the whole Troubadour corpus in order to render lemmatisation and morphosyntactic annotation possible for a larger corpus until exhaustivity is reached,
- to develop a cartography of Medieval Montpellier (from the Thalamus), and more generally a cartography of the Troubadour space in Europe,
- to ensure the alignment between the Medieval Occitan lexical resources and their modern and contemporary counterparts.

References

- [Caiti-Russo et al.2019] Gilda Caiti-Russo, Francesca Frontini, and Hervé Lieutard. 2019. Acolhir e Tornar – AcTo: Ressources numériques per l’occitan medieval [carnet de recherche]. <https://acto.hypotheses.org/>.
- [Carrasco et al.2014...] Raphaël Carrasco, Vincent Challet, Gilda Caiti-Russo, Stéphane Durand, Marc Conesa, Yves Mausen, Daniel Le Blévec, Chantal Wionet, and Florence Clavaud, editors. 2014/.... *Le «Petit Thalamus» de Montpellier: édition critique numérique du manuscrit AA9 des Archives municipales de Montpellier dit Le Petit Thalamus*. Université Paul Valéry Montpellier-III, Montpellier. <http://thalamus.huma-num.fr/>.
- [Clérice et al.2019] Thibault Clérice, Julien Pilla, and Jean-Baptiste-Camps. 2019. hipster-philology/pyrrha: 2.0.0. <https://doi.org/10.5281/zenodo.2541730>.
- [Crane2012] Gregory Crane. 2012. The Perseus Project. In *Leadership in Science and Technology: A Reference-Handbook*, pages 644–652. SAGE Publications, Thousand Oaks.
- [Guillot et al.2013] Céline Guillot, Sophie Prévost, and Alexei Lavrentiev. 2013. *Manuel de référence du jeu Cattex09*. École normale supérieure de Lyon, Lyon. Version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf.
- [Juršić et al.2010] Matjaz Juršić, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.
- [Manjavacas et al.2019] Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. *CoRR*, abs/1903.06939.
- [Passarotti et al.2019] Marco Passarotti, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, and Paolo Ruffolo. 2019. LiLa: Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin.
- [Pinche et al.2019] Ariane Pinche, Jean-Baptiste Camps, and Thibault Clérice. 2019. Stylometry for Noisy Medieval Data: Evaluating Paul Meyer’s Hagiographic Hypothesis. In *Digital Humanities Conference 2019 - DH2019*, Utrecht, Netherlands. ADHO and Utrecht University.
- [Prévost et al.2013] Sophie Prévost, Céline Guillot, Alexei Lavrentiev, and Serge Heiden. 2013. *Jeu d’étiquettes morphosyntaxiques CATTEX2009*. École normale supérieure de Lyon, Lyon. version 2.0. http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf.
- [Stempel et al.1996...] Wolf-Dieter Stempel, Maria Selig, Claudia Kraus, Renate Peter, and Monika Tausend. 1996/.... *Dictionnaire de l’occitan médiéval (DOM en ligne)*. Bayerische Akademie der Wissenschaften, Munich. <http://www.dom-en-ligne.de/>.
- [Wartburg1922 1967] Walther von Wartburg. 1922–1967. *Französisches Etymologisches Wörterbuch: eine Darstellung des galloromanischen Sprachschatzes*. ATILF, Leipzig. <https://apps.atilf.fr/lecteurFEW/>, *eFEW: FEW informatisé*, ed. Pascale Renders.
- [Wilhelm1970] James J. Wilhelm. 1970. *Seven Troubadours: Creators of Modern Verse*. Pennsylvania State University Press, University Park.

A Parsing Pipeline for Icelandic Based on the IcePaHC Corpus

Tinna Frímann Jökulsdóttir

University of Iceland
Reykjavík, Iceland
tinnafj@hi.is

Anton Karl Ingason

University of Iceland
Reykjavík, Iceland
antoni@hi.is

Einar Freyr Sigurðsson

The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
einar.freyr.sigurdsson@arnastofnun.is

Abstract

We describe a novel machine parsing pipeline that makes it straightforward to use the Berkeley parser to apply the annotation scheme of the IcePaHC corpus to any Icelandic plain text data. We crucially provide all the necessary scripts to convert the text into an appropriate input format for the Berkeley parser and clean up the output. The goal of this paper is thus not to dive into the theory of machine parsing but rather to provide convenient infrastructure that facilitates future work that requires the parsing of Icelandic text.

1 Introduction

Icelandic is a less-resourced language in the context of the CLARIN goals of fostering language resources and technology infrastructure; thus it is crucial to create further Icelandic resources that facilitate the development and use of Icelandic Language Technology for research as well as practical applications. Some efforts have been made in recent years to develop such resources and a notable example is the Icelandic Parsed Historical Corpus, IcePaHC (Wallenberg et al., 2011; Rögnvaldsson et al., 2012). The IcePaHC corpus contains one million running words of manually corrected phrase structure annotation. In this paper, we describe a novel parsing pipeline for Icelandic that makes crucial use of IcePaHC, thus making it straightforward for any future projects to take advantage of machine-annotation according to the IcePaHC annotation scheme. The pipeline is available on Github (<https://github.com/antonkarl/iceParsingPipeline>). This is ongoing work and we aim to package our solutions as CLARIN tools when they have reached a more mature state. To consider a concrete example from the corpus, the sentence below is taken from the 13th century manuscript called *Morkinskinna*.

- (1) *Par kom að þeim Danaher.*
there came to them Danish army.
'There, the Danish army came toward them.'

The corpus makes use of labeled bracketing similar to the Penn Treebank. The annotated version of the sentence is shown below.

```
( ( IP-MAT (ADVP-LOC (ADV þar-þar) )
  (VBDI kom-koma)
  (PP (P að-að)
    (NP (PRO-D þeim-það) )
    (NP-SBJ (NPR-N Danaher-danaher) ) ) (ID 1275.MORKIN.NAR-HIS, .522) )
```

One of the main reasons for constructing a parsed corpus at all is the ability to train an automatic parser in order to get access to fast machine annotation of the same type. While there are various interesting deep technical problems involved in machine parsing, the practical challenges of setting up convenient infrastructure for parsing are sometimes overlooked when discussing, say, the theoretical side of training parsers. Even if a parsed corpus is freely available, a researcher in the humanities or social sciences may

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

not have access to the technical background that is required to train a parser and set up a parsing pipeline. The current project aims to ameliorate this situation by developing a pipeline that takes Icelandic in plain text format, converts it to an appropriate input format for the Berkeley parser, parses the data using a pre-trained model that we provide, and cleans up the output. This, for example, allows anyone who has learned to use a treebank search program to apply this knowledge to their own data.

The paper is organized as follows: In Section 2, we introduce IcePaHC and related tools. In Section 3, we describe our matrix clause boundary detection system, and in Section 4 we discuss our use of the Berkeley parser. Section 5 offers information about some post-processing steps and Section 6 concludes.

2 The IcePaHC corpus and related tools

The Icelandic Parsed Historical Corpus (IcePaHC) is a dual purpose project. It is designed to serve both as a language technology tool and a syntactic research tool. It contains about one million words, fairly evenly distributed throughout the written history of the Icelandic language, from the 12th century to modern times. IcePaHC is released under a LGPL-license and it can be freely downloaded from http://www.linguist.is/icelandic_treebank/Download.

The annotation scheme, briefly demonstrated above, is based on the Penn Parsed Corpora of Historical English (Kroch and Taylor, 2000; Kroch et al., 2004). For most purposes, the English annotation guidelines are applied without modification to Icelandic and, in fact, the same search query can often be used for studying the same phenomenon in both languages. Some minor Icelandic-specific adjustments to the annotation scheme include a somewhat larger tagset and lemmatization; these reflect the morphological richness of Icelandic compared to English. The Icelandic guidelines also include a way to annotate non-nominative subjects, a well-known peculiarity of Icelandic syntax (Thráinsson, 2007).

The IcePaHC project has from the beginning emphasized a Free and Open Source (FOSS) policy. The development pipeline of the corpus itself consisted entirely of FOSS tools, including IceNLP for tagging and shallow parsing (Loftsson and Rögnvaldsson, 2007), Lemmald for lemmatization (Ingason, et al., 2008), CorpusSearch for structure building queries and quality checks (Randall, 2005), and Annotald for manual annotation (Beck et al., 2015). The project also gave rise to the Parsed Corpus Query Language (PaCQL), currently under development and available as a preview on www.treebankstudio.org. PaCQL has made it easier for scholars of the digital humanities to use IcePaHC. It should also be noted that another ongoing project aims to implement a conversion from IcePaHC annotation to Universal Dependencies, further expanding the technical ecosystem of the present development efforts.

3 Matrix clause boundary detection

A crucial step in parsing involves boundary detection of the segments that have a privileged status in the sense that they correspond to one tree in the annotation scheme. In the IcePaHC annotation scheme, this privileged unit is in most cases the matrix clause, meaning that we must have some means of detecting the boundaries of matrix clauses. In the original construction of the corpus, this step was carried out manually by research assistants, and the current project therefore needed to implement a software solution for this task. We divided the task into two subtasks, (i) punctuation-based sentence boundary detection and (ii) conjunction-based matrix clause boundary detection. For both steps, we configured a feature extractor for potential boundaries and used IcePaHC to train an implementation of the Averaged Perceptron classifier (Freund and Schapire, 1999) by Kyle Gorman (Gorman, 2019) to detect actual boundaries.

For the punctuation-based sentence boundary detection we used the Python package Detector Morse (Gorman, 2019) and the feature extractor that comes with this package. The machine-learning classification algorithm, the Averaged Perceptron, is loaded from the `nlp` Python package, also authored by Gorman. Detector Morse was originally developed for English but its design carries over straightforwardly to languages that use a similar alphabet and similar conventions for spelling and punctuation. As expected, training Detector Morse in its default configuration by using the sentence splits in IcePaHC as a gold standard gave good results without the need for any Icelandic-specific adjustments.

The output from Detector Morse is subsequently fed into a conjunction-based matrix clause boundary detection system. The design of our matrix clause splitter is based on Detector Morse, and it uses the

same implementation of the Averaged Perceptron, but in this case it is not possible to achieve good results without the ability to detect language-specific morphosyntactic properties of the context. Therefore, we developed our own feature extractor for this task.

The conjunctions *og* ‘and’, *en* ‘but’ and *eða* ‘or’ are considered potential matrix clause boundaries whenever they are attested. In each case, the classifier must determine if it is an actual matrix clause boundary, like *and* in (2) or a different use of the relevant conjunction, like *and* in (3).

(2) John walked to the store **and** Mary smiled when she saw him.

(3) John **and** Mary walked to the store.

The feature extractor considers two words preceding the conjunction and two words following the conjunction and the pattern it returns is sensitive to whether any of these words, in this relative position to the conjunction, is potentially a morphosyntactic indicator. The indicators in question are: comma, finite verb, non-finite verb, a word in the nominative case, a word that has any non-nominative (oblique) case value, i.e. accusative, dative or genitive case.

While this setup works relatively well, the focus of the current phase of our project is simply to get something working that can be used for practical applications, rather than optimizing individual steps. Our morphosyntactic indicators are chosen based on our experience of Icelandic syntax, but we are confident that the feature extraction can be improved and this will be done in future work along with proper evaluation. For now, the matrix clause splitter works well enough for most detected matrix clauses to be a suitable input to the Berkeley Parser as trained on IcePaHC.

4 Training the Berkeley Parser

We chose the Berkeley Parser (Petrov et al., 2006) for our work because its split-merge algorithm is known to yield accurate results, it is relatively simple to use for data that are already in a labeled bracketing file format, and there exists a version of it that runs fast on massively multi-core GPU cards, i.e., its Puck implementation (Canny et al., 2013; Hall et al., 2014). The ability to run the parser on GPU’s is particularly important for large data sets and High-Performance Computing Clusters (HPCC) that emphasize GPU-computing are increasingly becoming a part of the technology infrastructure that research universities and organizations are continually expanding. We use the Berkeley Parser for Part-of-Speech tagging as well as for parsing phrase structure.

The training data we used was the full IcePaHC corpus. Although the parser assumes labeled bracketing, we did need to make a few minor adjustments to the file format to get the training phase to run smoothly, and this is important in the context of the present paper because it exemplifies how practice is more complicated than theory for a language technology task like machine parsing. In theory, having a parsed corpus and a freely available implementation of some trainable parser is enough to yield machine parsing for the annotation scheme in question, but in practice it takes quite a bit of technical work to set up the process. For researchers in the humanities and social sciences who may just want access to the annotation without having to put unreasonable effort into getting the system to work, having access to a pre-configured parsing pipeline can make an important difference in getting the results they want.

5 Post-processing and cleanup

Our pipeline also includes scripts that take the output of the Berkeley parser and make some minor adjustments to it. While these steps do not change the information that is included in the output from the parser, they make its format more similar to what scholars who study historical syntax are used to.

The community of researchers who study historical syntax using treebanks and quantitative methods has by now become very familiar with the files that are used for the raw data of the Penn Parsed Corpora of Historical English because these corpora are frequently used and other corpora that have been developed for the same group of users have adopted this format. This format is, of course, machine-readable, but as the conventions used for indenting etc. have become somewhat standard for this subset of treebanks, it is, for an experienced user, also conveniently human-readable. While such formatting

nuances might be considered an unimportant detail, we believe that it will make our parsing pipeline more pleasing to use for the users that are most likely to make it a part of their workflow.

6 Summary and future work

We have described our parsing pipeline for Icelandic that takes plain text input and yields output that is machine-annotated according to the annotation scheme of the IcePaHC treebank. This includes pre-processing, parsing using a pre-trained model of the Berkeley-parser, and formatting and cleanup of the output. We make these steps easily executable by any future project that may benefit from parsing Icelandic. While our parsing pipeline is already available and useful, much remains to be done. The configuration that we use for individual steps such as training the matrix clause splitter and the parser will be improved in future work in order to yield even better results and proper evaluation will be an essential ingredient of any iterative improvements. At this point, we focus on a pipeline that can be used for further development, hence evaluation of different parser configurations remains a future task. With our pipeline in place, we have also started work on a Machine-parsed IcePaHC (MICEPaHC), a corpus that does not have the manual corrections of IcePaHC, but can grow much faster in size because all of the annotation is carried out by computers. We aim to release the first version of MICEPaHC in the near future, both in terms of freely available raw data and as a search option on treebankstudio.org.

References

- Beck, Jana, Aaron Ecat, and Anton Karl Ingason. 2015. Annotald. Version 1.3.7.
- Canny, John, David Hall, and Dan Klein. 2013. A multi-Teraflop Constituency Parser using GPUs. *Proceedings of Empirical Methods in Natural Language Processing*, pp. 1898–1907.
- Freund, Yoav, and Robert E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning* 37.3 (1999): 277–296.
- Gorman, Kyle. 2019. Detector Morse. A Python Package. Version 0.4.1.
- Hall, David, Taylor Berg-Kirkpatrick, and Dan Klein. 2014. Sparser, Better, Faster GPU Parsing. *Proceedings of ACL 2014*, pp. 208–217.
- Ingason, Anton Karl, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). *Proceedings of GoTAL*, pp. 205–216. Springer, Berlin, Heidelberg.
- Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. 2004. Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME). Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (<http://www.ling.upenn.edu/ppche-release-2016/PPCEME-RELEASE-3>).
- Kroch, Anthony, and Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English (PPCME2). Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (<http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4>).
- Loftsson, Hrafn, and Eiríkur Rögnvaldsson. 2007. IceNLP: A natural language processing toolkit for Icelandic. *Proceedings of Eighth Annual Conference of the International Speech Communication Association*.
- Petrov, Slav, Leon Barrett, Romain Thibaux and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. *Proceedings of COLING-ACL 2006*, pp. 433–440.
- Randall, Beth. 2005. CorpusSearch 2. User's manual.
- Rögnvaldsson, Eiríkur, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). *Proceedings of LREC'12*, pp. 1977–1984.
- Thráinsson, Höskuldur. 2007. *The Syntax of Icelandic*. Cambridge University Press, Cambridge.
- Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. The Icelandic Parsed Historical Corpus, version 0.9. 1 million words.

Optimizing Interoperability of Language Resources with the Upcoming IIF AV Specifications

Jochen Graf

University of Cologne
jochen.graf@uni-koeln.de

Felix Rau

University of Cologne
f.rau@uni-koeln.de

Jonathan Blumtritt

University of Cologne
jonathan.blumtritt@uni-koeln.de

Abstract

In our presentation, we discuss how the upcoming IIF AV specifications could contribute to interoperability of annotated language resources in the CLARIN infrastructure. After some short notes about IIF, we provide a comparison between the concepts of the IIF specifications and the ELAN annotation format. The final section introduces our experimental *Media API* that intends to optimize interoperability.

1 Introduction

The International Image Interoperability Framework (IIF) (Snydman et al., 2015) is a technology agnostic standardisation for dissemination of web based images. It is driven forward by a large community of cultural heritage institutions. The original motivation behind IIF is to optimize interoperability such that annotated image resources available at one institution can be reused by tools and services at other institutions. A side effect is that the framework facilitates implementation of web based image and annotation clients. With IIF, clients can rely on well defined, feature rich, and stable application programming interfaces.

The IIF Image API (Consortium, 2017a) and the IIF Presentation API (Consortium, 2017b) build the core APIs of the framework. The IIF Image API defines a set of low-level image manipulation requests, e.g., for image cropping, rotation, or format conversion. These low-level requests enable higher level features relevant for interoperability: for example, persistent web references not only to whole images but also to image details. The main idea behind the IIF Presentation API is the so called *Canvas*¹. *Canvas* represents a 2D coordinate space, where the target image(s) to be annotated and the annotations itself are organized together. The strength of the *Canvas* lies in its abstraction. The 2D canvas can be replaced by a canvas timeline for AV annotations with only small modifications on the specifications necessary.

Once an image-centric framework, IIF is currently extended to other resource types from the cultural heritage domain, especially too for AV resources. Since 2016, the IIF AV Technical Specification Group (Consortium,) develops a new *AV Content API* mirroring the IIF Image API in function and refines the IIF Presentation API in order to make it equally useful for image, audio, and video annotation.

In the following, we aim to show how the upcoming IIF AV specifications could contribute to the interoperability of language resources and to the development of web based AV annotation players - and the other way round: we aim to show that the ELAN annotation format forms an interesting case study to further develop the IIF AV specifications.

2 ELAN IIF AV Case Study

ELAN (Wittenburg et al., 2006) is a desktop annotation software for multimodality research. It is, among others, central to the research in the communities represented in the CLARIN-D working group *Linguistic Fieldwork, Ethnology, and Language Typology*. The tool produces time-aligned annotations in ELAN Annotation Format (EAF) (The Language Archive,), an open XML format. The ELAN tool is suitable for linguistic research since it does not only allow simple transcription of audio data but discipline-specific, time-aligned annotations up to the level of syllables and phonemes. A large number of EAF

¹<https://iif.io/api/presentation/2.1/#canvas>

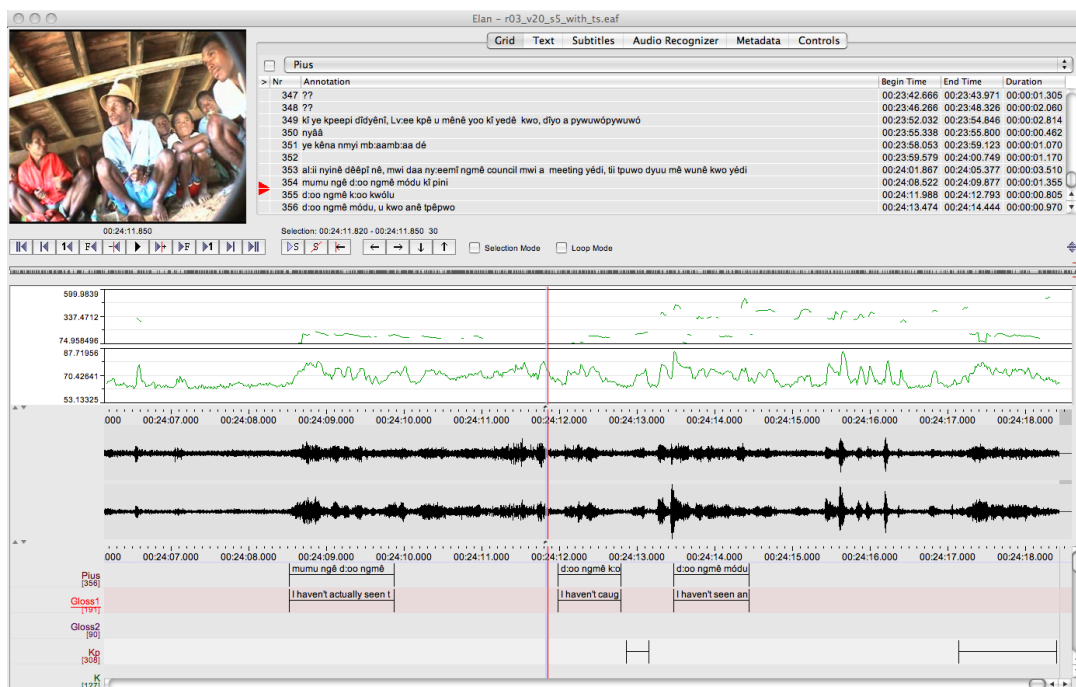


Figure 1: ELAN Main Window

documents has become available in numerous language archives in recent years. To the best of our knowledge, there are only few archives that implement web based viewers for EAF Annotations in an adequate way (Berck and Russel, 2006)(Schroeter and Thieberger, 2006)(Sjölander and Beskow, 2000). As a result, the need for web interoperability of AV annotations should be as natural as for metadata (Freire et al., 2017).

Figure 1 (for Psycholinguistics,) shows the ELAN main window. Directly above and below the zoomable timeline in the middle, there are visual representations of the video's audio channels (intonation, waveform) - helpful tools providing researchers with an overview for an inherently time-bound and transient type of data. At the bottom, there are time aligned annotations grouped by layers shown in different colors. If one plays the referenced video file, the timeline and its attached intonation, waveform, and annotations are presented in form of a concurrent display with high time accuracy.

Having a deeper look into the ELAN annotation format, there appears much more complexity on the annotation level than visible on the user interface. The annotation format does not only support time aligned annotations² but also annotation references³. Annotation references have no direct timeline linkage but are linked to a parent annotation⁴. Additionally, there exist annotation types that can either subdivide the time range of a parent annotation having an own fixed start and end point in time⁵, or types that dynamically divide parent annotations into a defined number of parts with equal length and without gaps⁶.

When comparing the ELAN annotation format with the ongoing work done by the IIF AV Technical Specification Group, we identify overall accordance in respect to the way annotations are structured and

²<eaf:ALIGNABLE_ANNOTATION/>

³<eaf:REF_ANNOTATION/>

⁴Symbolic_Association

⁵Time_Subdivision and Symbolic_Subdivision

⁶Included_In

grouped in lists and layers, enriched with different types of metadata, and linked to (parts of) media files. Those concepts can be easily mapped in both directions. We currently identify two differences, though:

Difference 1. The ongoing IIF AV Content API specification does not yet propose the generation of visual representations of audio data such as spectrums or waveforms, although this is a useful utility for linguistic research. In the context of IIF, it seems obvious to provide such visual representations in the form of images, respectively, with image tiles. A image tile of 500x25 pixels would contain the spectrum of a audio's time section with 10 seconds in length, for example. If a number of spectrum tiles is seamlessly strung together, concurrent display and deep zooming of time aligned annotations becomes possible in the browser as with ELAN, even for very large audio files.

Difference 2. There are many different types of annotations supported by the ELAN annotation format that, in their discipline-specific variety, seem to lay beyond the expressiveness of IIF annotations. Since we could not find a convincing mapping, our current approach is to transform linguistic annotation references into standard, time aligned annotations accepting information loss. Since it is not our aim to provide a web version of the ELAN annotation tool, but only a player for presentation of AV annotations, the loss of information seems acceptable in regard to the increased interoperability.

3 Media API

3.1 Requirements Analysis

The requirements analysis for our experimental Media API started with a description of the *ELAN IIF AV Case Study* in order to tie our experiment to a large real-world AV annotation dataset. Our Media API in any case should follow the IIF AV specifications in the way that it mirrors the IIF Image API in function: the proposed API should support all common transformations on AV media (cropping, format conversion, etc.) in a simple way as is the case for images and the IIF Image API. Since our case study has shown that the scientific practice of linguistic annotation is well supported by visual representations of audio data, i.e., by spectrum or waveform image tiles, we like to propose that an ideal Media API would not only mirror the image API in function but would ideally be a superset of the IIF Image API. In summary, we expect the Media API to cover the following function areas with at least the functionality mentioned in brackets:

Requirement A: common media transformations (format conversion, compression)

Requirement B: audio/video specific transformations (time cropping)

Requirement C: video/image specific transformations (cropping, scaling, rotating, color filtering)

Requirement D: audio to image transformations (spectrum and waveform extraction)

3.2 Derivation of the Media API from the IIF Image API

The IIF Image API defines five request parameters for image transformation as summarized below. According to the IIF specifications, the parameters are processed in the order they are arranged in the URI from left to right: first, a rectangular portion of the input image is cropped, then the image is scaled, and so on.

Canonical URI Syntax of the IIF Image API:

```
.../{region}/{size}/{rotation}/{quality}.{format}
```

Request Parameters of the IIF Image API:

{region}	Defines the rectangular portion of the full image to be returned.
{size}	Determines the dimensions to which the extracted region is to be scaled.
{rotation}	Specifies mirroring and rotation.
{quality}	Determines whether the image is delivered in color, grayscale or black and white.
{format}	Format of the returned image.

Based on this API, we implemented our experimental *Media API* that allows to display linguistic annotations, visualizations of the audio signal as well as playback of the audio-visual data itself. The implementation prioritises interoperability with the IIF AV specifications. For our API, we adopted the ongoing IIF AV specifications and extended the concepts to fit time-aligned linguistic annotations.

Canonical URI Syntax of the Media API:

```
.../{section}/{region}/{size}/{rotation}/{filter}/{quality}.{format}
```

Request Parameters of the experimental Media API:

{section}	Defines the time portion of the full audio or video file to be returned.
{region}	Defines the rectangular portion of the full image or video to be returned.
{size}	Scales an image or video to a specific size.
{rotation}	Specifies mirroring and rotation for a image or video file.
{filter}	Applies filters to the input media file (waveform, spectrum, color, gray, bitonal, none).
{quality}	Defines the compression rate / quality scale of the returned media file (high, medium, low).
{format}	Format of the returned media file.

Our experimental media API, compared to the IIF Image API, contains three AV related extensions:

Extension 1. In times of mobile devices used in low bandwidth networks, it seems desirable to offer audio and video data in different quality scales. For this purpose, we decided to reinterpret the *{quality}* parameter directly before the *{format}* ending due to its purely image related meaning (color, grayscale, black and white). *{quality}* in our media API does not refer to the visual quality of the returned image but to the technical quality scale of the media file, where *high*, respectively *default* return the image, audio or video bitstream as is, *medium* and *low* return an increasingly compressed version of the input file with possibly human perceivable quality loss. Extension 1 fulfills requirement A.

Extension 2. The original *color*, *grayscale*, *black and white* functions are still there but are moved to the *{filter}* parameter. Together with the *{region}/{size}/{rotation}* URI part, requirement C is fulfilled and the API parameters together form a superset of the IIF Image API. "Grayscale filter" or "bitonal filter" seem still acceptable names for the respective functions. The *{filter}* parameter introduces flexibility for different media types: if the input file is a audio or video, one can apply a spectrum or waveform filter here. A spectrum image of 500x25 pixels in PNG format calculated from a 10 seconds audio section can be requested as follows:

```
.../0,10/full/500,25/0/spectrum/default.png
```

Extension 2 fulfills requirement D.

Extension 3. Finally, a *{section}* parameter is put in front of all other parameters in order to allow cropping of time sections of audio and video files. This fulfills requirement B.

4 Conclusion

Adopting the ongoing IIF AV specifications and extending its concepts to fit time-aligned linguistic annotations is showing promising results. Our current work concentrates on writing down a detailed technical documentation of our case study and on developing a prototype of our Media API - with the intent to report back our results to the IIF AV Technical Specification Group. In order to achieve full interoperability within the CLARIN infrastructure, other issues have to be addressed, though: foremost, the issue of authentication and authorization of REST APIs in SAML-based authentication and authorization infrastructures needs further attention.

References

- Peter Berck and Albert Russel. 2006. ANNEX a web-based framework for exploiting annotated media resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).
- IIF Consortium. IIF AV technical specification group folder. <https://drive.google.com/drive/folders/0B8SS5OUXWs4GZ0ZfbEhIc1hzb0k>. [Online; accessed 2019-04-09].

- IIIF Consortium. 2017a. IIIF image API 2.1.1. <https://iiif.io/api/image/2.1/>. [Online; accessed 2019-04-09].
- IIIF Consortium. 2017b. IIIF presentation API 2.1.1. <https://iiif.io/api/presentation/2.1/>. [Online; accessed 2019-04-09].
- Max Planck Institute for Psycholinguistics. ELAN main window. <https://tla.mpi.nl/tla-news/annex-and-elan-a-comparison>. [Online; accessed 2019-04-09].
- Nuno Freire, Glen Robson, John B. Howard, Hugo Manguinhas, and Antoine Isaac. 2017. Metadata aggregation: Assessing the application of IIIF and sitemaps within cultural heritage. In Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis, and Ioannis Karydis, editors, *Research and Advanced Technology for Digital Libraries*, pages 220–232, Cham. Springer International Publishing.
- Ronald Schroeter and Nicholas Thieberger. 2006. EOPAS, the EthnoER online representation of interlinear text. In *Sustainable Data from Digital Fieldwork. Proceedings of the conference held at the University of Sydney, 4-6 December 2006*. Sydney University Press.
- Kåre Sjölander and Jonas Beskow. 2000. Wavesurfer – an open source speech tool. In *Sixth International Conference on Spoken Language Processing*.
- Stuart Snyderman, Robert Sanderson, and Tom Cramer. 2015. The international image interoperability framework (IIIF): A community & technology approach for web-based images. *Archiving Conference*, 2015(1):16–21.
- Max Planck Institute for Psycholinguistics The Language Archive. ELAN annotation format. http://www.mpi.nl/tools/elan/EAF_Annotation_Format_3.0_and_ELAN.pdf. [Online; accessed 2019-04-09].
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

SpeCT 2.0 – Speech Corpus Toolkit for Praat

Mietta Lennes

Department of Digital Humanities / FIN-CLARIN

University of Helsinki, Finland

mietta.lennes@helsinki.fi

Abstract

Praat (Boersma and Weenink, 2019) is a versatile, open-source platform that provides a multitude of features for annotating, processing, analyzing and manipulating speech and audio data. By using the built-in scripting language, Praat can be easily extended and adjusted for different purposes while reducing manual work. The Speech Corpus Toolkit for Praat (SpeCT) is a collection of Praat scripts that can be used to perform various small tasks when building, processing and analyzing a speech corpus. SpeCT can help both beginners and advanced users solve some common issues in, e.g., semi-automatic annotation or speech corpus management. This work describes some of the general functionalities in SpeCT. A selection of the scripts will also be made available via the Mylly service at the Language Bank of Finland, maintained by FIN-CLARIN.

1 Introduction

Praat (Boersma and Weenink, 2019) is a well-established and popular speech analysis platform not only for speech researchers but also for many others working with audio signals. Praat is based on open source code and the package is maintained on GitHub¹. Praat is available for Windows, Mac and Unix/Linux and it can be downloaded and used free of charge.

Originally, Praat is “a system for doing phonetics by computer” (Boersma and Weenink, 2019). Indeed, it provides a huge number of useful functionalities for analyzing, visualizing and annotating speech signals. However, Praat is much more than a speech transcription workbench. For instance, it also provides many statistical analyses, an articulatory speech synthesizer (described in (Boersma, 1998)), features for doing source-filter synthesis, an editor for manipulating pitch and durations within sound files, and a system for creating graphics for publications.

Importantly, Praat also includes a full-scale scripting language that makes use of all the high-level menu commands of the graphical user interface in addition to general programming syntax. Praat scripts can be run in the graphical user interface or in batch mode, and they can be called by other programs. Nearly identical features, displays and file formats are available in all operating systems. Thus, Praat represents a speech data ecosystem that can be readily combined with other digital research environments and analysis pipelines.

*The Speech Corpus Toolkit for Praat (SpeCT)*² is a collection of Praat scripts that can be used to perform various small tasks when building, processing and analyzing a speech corpus. Earlier versions of many of the scripts have been available online since 2001. The repository was renamed as “the Speech Corpus Toolkit for Praat” in 2011 and moved to GitHub in 2017 (Lennes, 2017). From the start, the scripts have been well commented. Instructions for particular tasks are offered on supplementary web pages³ in order to reduce the need for email support, since only one person has been responsible for maintaining the scripts. However, during the past few years, there has been an increasing need for updating the collection, for fixing bugs and for making the scripts more consistent and interoperable.

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

¹Praat repository on GitHub: <https://github.com/praat>

²The Speech Corpus Toolkit for Praat on GitHub: <https://github.com/lennes/spect>

³SpeCT documentation: <https://lennes.github.io/spect/>

SpeCT was originally created in order to provide accessible and lightweight solutions to some frequently occurring practical problems and tasks in speech research, especially bearing in mind the “less technical” researchers and students. In addition to SpeCT, there currently exists a wealth of other online resources containing various kinds of Praat scripts and plugins that could potentially be used as components in customized workflows. However, for more experienced programmers it is useful to look into Python modules and libraries related to Praat, e.g., `pypmi` (Lubbers and Torreira, 2013 2018), and/or interfaces between Praat and statistical analysis software, e.g., `praatR` (Albin, 2014).

2 Speech corpus workflow

After recording some audio material for a speech corpus, researchers, research assistants or students tend to face many practical problems. SpeCT can provide at least partial solutions to some typical issues in organizing and handling speech corpora.

2.1 Slicing sound files

Sound files often require some trimming and sometimes they need to be sliced into shorter excerpts. SpeCT contains scripts with different options for exporting annotated segments into individual sound files. The user can mark the portions either semi-automatically (by first letting the computer mark the quiet portions as silences) and/or manually according to visual and auditory inspection. This process reduces the manual work load, and the resulting files can be automatically named in a systematic way, thus minimizing the risk for errors.

2.2 Pause detection and semi-automatic alignment of transcripts

At some point, the original sound files usually need to be transcribed and/or annotated with other necessary details so as to eventually allow for the efficient searching and analysis of the complete corpus. However, manual annotation is typically very time-consuming, and for many languages and especially for casual conversational speech, fully automatic speech recognition tools might not produce satisfactory results. Therefore, any kind of semi-automatic aid may turn out to be useful during the annotation stage.

In SpeCT, there are scripts for detecting relatively silent portions of the sound signal (“pauses”), and a similar feature is also built in Praat. Moreover, in case the user already has a text file containing a transcript of the conversation, it is possible to use a script in SpeCT for semi-automatically aligning the transcribed utterances or turns with the audio file. The script works by suggesting the start and end boundaries for one utterance or turn at a time, and the user needs to modify the suggested boundary locations if required. The suggestions are based on detecting silent portions as well as the number of characters in the transcribed line of text. Each speaker’s turns are marked in a separate annotation tier. Naturally, the method is quite crude and intervention is required from the user, but since a language-independent heuristic is used, the script has turned out to be quite helpful in some cases.

2.3 Search and look-up from a number of corpus files

Sometimes systematic transcription errors are spotted while the annotation project is still ongoing. Using a Praat script, it is possible to look up a given portion of a specific annotated sound file in a large corpus, without remembering the name of the file. If certain transcriptions need to be changed in the entire corpus, the replacement can be done automatically by using a script.

If the annotation project is large, another script can be used in order to obtain an overview of which files have been annotated already, what tiers they include, how many annotation items they contain, or which files lack particular annotation tiers.

2.4 Measuring and extracting samples for specific purposes

In the analysis stage, the researcher may require a tool for measuring and collecting acoustic-phonetic parameters (e.g., durations, pitch, formant frequencies etc.) or complex linguistic properties that may combine time-synchronous data from several annotation tiers as well as the original audio signal in the annotated corpus. The analysis procedure is likely to be quite specific for each research question,

so a general-purpose search tool may not suffice. Several different scripts are available for exporting measurements and other data from Praat to spreadsheet documents or to other interoperable data formats, or for plotting measured data in figures.

Many more speech corpus issues can be addressed by using the Praat scripts already included in SpeCT. It is also relatively straightforward to adapt the scripts in order to match other similar purposes.

3 Future work

During 2019, a second generation of SpeCT will be created by modifying and updating the existing Praat scripts according to more consistent design principles. The script collection will continue to be publicly available via GitHub. In addition, as part of the updating process of SpeCT, a sample of the Praat scripts will also be installed for test use in the researchers' toolbox called *Mylly*, 'the Mill' (Lennes et al., 2017), one of the services available at the Language Bank of Finland.⁴ In *Mylly*, the user can log in to the online service, upload files to a personal workspace and process them with various tools by simply selecting the desired commands from the menu. *Mylly* allows users to build and save workflows for reuse.

For SpeCT, the general development goal is to ensure that each script can be conveniently used either in isolation or as part of a longer workflow that might include processing steps outside the Praat environment. In order to fulfil the development goal, each script should be self-contained and well instructed. For input and output, the scripts in SpeCT should rely on native Praat objects or, if necessary, on simple and generic data formats. By complying to these principles, it will be even easier to modify the scripts or to combine them with other similar resources. In the conference, we will demonstrate some of the main tools provided by SpeCT for the different stages of the speech corpus processing workflow.

References

- Aaron Albin. 2014. PraatR: An architecture for controlling the phonetics software "Praat" with the R programming language. *Journal of the Acoustical Society of America*, 135(4):2198.
- Paul Boersma and David Weenink. 2019. Praat: doing phonetics by computer (Version 6.1.03). [Computer program]. <https://www.praat.org/>.
- Paul Boersma. 1998. *Functional phonology. Formalizing the interactions between articulatory and perceptual drives*. Ph.D. thesis, University of Amsterdam.
- Mietta Lennes, Jussi Piitulainen, and Martin Matthiesen. 2017. Mylly — The Mill: A new platform for processing speech and text corpora easily and efficiently. In *Interspeech 2017: Show & Tell Contribution, Stockholm, Sweden*, pages 829–830.
- Mietta Lennes. 2017. SpeCT — The Speech Corpus Toolkit for Praat (v1.0.0). First release on GitHub. Zenodo. <http://doi.org/10.5281/zenodo.375923>.
- Mart Lubbers and Francisco Torreira. 2013-2018. *pypi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files*. <https://pypi.python.org/pypi/pypi-ling>. Version 1.69.

⁴Kielipankki, The Language Bank of Finland (<https://www.kielipankki.fi>) is a collection of services maintained by FIN-CLARIN.

CLARIN-IT and the Definition of a Digital Critical Edition for Ancient Greek Poetry: a New Project for Ancient Fragmentary Texts with a Complex Tradition

Anika Nicolosi
Department DUSIC
University of Parma, Italy
anika.nicolosi@unipr.it

Monica Monachini
ILC - CNR
Pisa, Italy
monica.monachini@ilc.cnr.it

Beatrice Nava
Alma Mater Studiorum
University of Bologna, Italy
beatrice.nava2@unibo.it

Abstract

Ancient Greek studies, and Classics in general, is a perfect field to demonstrate how Digital Humanities could become the humanist way of building models for complex realities, analysing them with computational methods and communicating the results to a broader public. Ancient texts have a complex tradition, which includes many witnesses (texts that handed down another texts) and different typology of supports (papyri, manuscripts and also epigraphy). These texts are fundamental for our cultural Heritage, since they are the basis of all European Literatures, and it is crucial to spread their knowledge, in a reliable and easy way.

Our project on ancient Greek fragmentary poetry (DEA - *Digital Edition of Archilochus: New models and tools for authoring, editing and indexing an ancient Greek fragmentary author*) develops and grows out of existing experiences and try to define a new digital and critical edition which includes the use of Semantic Web and Linked Open Data. Our goal is to provide a complete and reliable tool for scholars, suitable for critical study in the field, and also user-friendly and useful for non-specialist users. The project represents one of the attempts within the context of CLARIN-IT to contribute to the wider impact of CLARIN on the specific Italian community interested to Digital Classics and may improve services in fostering new (and sustaining existing) knowledge in SSH digital research.

1 State of the art: DH and Classics

Ancient Greek studies, and Classics in general, is a perfect field to demonstrate how Digital Humanities could become the humanist way of building complex models of complex realities, analysing them with computational methods and communicating the results to a broader public. The Digital Classics have undergone a great development, starting from the last ten years of the XXth century, and today we have many sources available and refined tools – for example the main tools in the field as Thesaurus Linguae Graecae (TLG) and Perseus Digital Library (PDL). There are also other very important projects as LOFT, DCLP (and Trismegistos in general), MP³- CEDOPAL, Musisqueoquoque, Pinakes, and bibliographical tools as APh, but they are not complete and they do not interact with each other, so they can't replace the consultation of paper editions. Moreover, they are known only from specialist users.

The treatment of literary texts currently does not correspond to scholars' expectations. These are texts with a complex tradition, which includes many witnesses (texts that handed down other texts) and different typology of supports (papyri, manuscripts and also inscriptions). To have a complete knowledge, it is not enough to provide a text, we need much more information, that the scholar usually obtains comparing several paper editions, lexica and / or more digital tools and imagines.

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

2 Ancient Greek Poetry and Digital Edition: texts with a complex tradition

DH may improve and realize something new. It is not enough today only to describe materials, and/or to give information, and/or to make available texts and/or to analyse them, etc. It is necessary to integrate all these functions in a unique workbench where researcher can find all she/he needs. And to do this is essential to focus on a specific element. We think that ancient Greek literature, thanks to its potentialities and peculiarities, can be the right test.

Thanks to modern science we can preserve and handed down ancient Greek text to the next generation (not only among specialist in the field), exactly what ancient scholars did many centuries ago at the Alexandrian Library (3rd c. BCE). Fragmentary ancient Greek poetry is very different from other literary texts. In fact, its tradition is more complex since it has different kinds of sources (manuscripts, papyrus, epigraphy) with variants and lacunae. It is not enough to provide a single text chosen at random; it is necessary to carry out a complete revision of the texts, according to their updated critical edition, and to take into account all the textual proposals made by scholars. In this way, studying previous editions and secondary literature, it is possible to enrich the new edition with new hypotheses.¹

The digital medium allows us to manage much more data than we can do in a paper edition, in which the apparatus must often be proportionate to the size of the page and therefore often provides only the main proposals. We can easily manage and store in a single place all the hypotheses of previous editors and the additional useful information linked to the edition; therefore, we can facilitate a new philological approach offering all the interpretations of previous editors through a single resource, with the addition of new scholars' hypotheses.

In the same time, applying NLP to a fragmented tradition can be particularly challenging because we have multiple options for text reconstruction. Automatic linguistic analyses of the whole corpus can indeed not only support new readings and interpretations, but also lead us to greater certainty as regards text corrections, integrations and authorship. Moreover, making the data researchable and interoperable, we can facilitate a systematic production of specific lexica based on the ancient fragments and became a stable and immutable sample for automatic translation experiments.

2.1 DEA Project: A new Digital and Critical Edition

DEA, which stands for *Digital Edition of Archilochus: New models and tools for authoring, editing and indexing an ancient Greek fragmentary author*, is a project by principal investigator Anika Nicolosi (University of Parma) done in collaboration with ILC-CNR of Pisa and CLARIN-IT. Our project develops and grows out of existing experiences and try to define a new and complete digital edition of Archilochus' fragments, which includes the use of Semantic Web and Linked Open Data.² We have around 300 fragmentary poems by this important author, who lived in the 7th century BCE and who was closely related to Homer. Some fragments have only recently been published, however, what is currently lacking is a complete on-line critical edition of his works³. The main objective of the project is to provide scientifically reliable texts, with critical apparatuses, commentaries and translations, and to make available an online and easily accessible augmented corpus of ancient Greek fragmentary literature. For this reason, it is important to find an easy and effective way of managing all these data, also respecting the dictates and the needs of the philological tradition.

To reach our objectives and to set up the project correctly, we have to test and find solutions to make all data researchable and easily interoperable. The first step is the TEI transcription and the annotation of the text. Thanks to the Semantic web and Linked Open Data we can also enrich our knowledge and produce a much more complete information. For example, we can improve our edition with geo-spatial references, using for instance the Pleiades gazetteer of ancient places. But it is also necessary a more detailed annotation level that configures the text in its complexity and make visible gaps, supplements, and doubtful readings (see Figure 1).⁴

¹ See (Nicolosi 2015).

² See (Monachini, Khan, Frontini and Nicolosi 2018) and (Brando, Frontini and Ganascia 2016).

³ The work starts from (Nicolosi 2013) and (Nicolosi 2017).

⁴ The text's reference for these exempla is Archil. fr. 23 W.², studied by Beatrice Nava during her BA level thesis in Classics, a.a. 2013/2014, Tutor Prof.ssa A. Nicolosi, at the University of Parma.

```

<1
n="6">[.].<unclear>β<unclear>α.....<unclear>δ<unclear>ε.<unclear>ή<unclear>μειβόμ<supplie
d resp="#Lobel">ην </supplied></1>
  <1 n="7">"γόνα<supplied resp="#West">ι</supplied>, φάτιν μὲν τὴν <unclear
resp="#West">πρ<unclear>ὸς ἀνθρώπ<unclear>ω<unclear><supplied resp="#West">ν
καὶν</supplied></1>
  <1 n="8">μὴ τετραμῆνης μὴ<unclear>δ<unclear>έν. ἀμφὶ δ'εὐ<unclear>φ<unclear><supplied
resp="#West">φρόνητι,</supplied></1>
  <1 n="9">ἔμοι μελήσει. <supplied resp="#West">θυμὸν ἰλ<unclear>α<unclear>ον
τίθεο.</supplied></1>
  <1 n="10">ἐς τοῦτο δὴ τοι τῆς ἀνολβίης δοκ<supplied resp="#West">έω</supplied></1>

```

Figure 1

We are currently studying a pilot set of Archilochus' papyrus texts, in which there are several typologies of textual problems, as gaps and supplements. Our goal at this stage is to find appropriate solutions to manage the critical apparatus and witnesses with TEI encoding. The current hypothesis has three different annotation levels, divided along different typologies of witnesses and different text's references.

1. The tag <listWit> includes:
 - a) ancient witness, associated with an xml: id and some essential information (<msDesc>)
 - b) modern editions, each associated with an xml: id and the complete bibliographic reference (<bibl>)
2. The tag <listBibl> includes bibliographic references, associated with an xml: id of:
 - a) secondary literature
 - b) ancient texts (not indirect witnesses, but used for conjecture)

The coding of the apparatus is made with the parallel segmentation method. In the apparatus (<app>) we can differentiate between, for example:

- a) Lesson of the reference text (e.g. Nicolosi), for example with the tag <lem>ἰλ<unclear>α<unclear>ον</lem>
- b) Reading proposed in a modern edition (related to listWit), for example <rdg wit="#Latte_1955">ἰλ<unclear>ε<unclear>ον</rdg>
- c) Reading from secondary literature (reference refers to listBibl), for example <rdg wit="#Bossi_1990">example</rdg>
- d) Hypothesis of a modern editor supported by an ancient text, indicated with @source; here the reference refers to listBibl, for example <rdg wit="#Adrados_1990" source="#Exemplum">example</rdg>

To standardize the model, we sum up all these data and we created a TEIheader that could be the guideline to replace the approach to ancient text (Figure 2 and 3).

Typology	Mark
Title	<titleStmb><title>Archilochus, fr. 23 W.<chi rend="apex">2</chi>, digital edition</title>
Author (and other responsibilities)	<respStmb> <resp>Encoding (or other responsibilities)</resp> <persName xml:id="Iniziali nome puntate">Name and Surname</persName> </respStmb> </titleStmb>
Digital Edition	<publicationStmb> <publisher>D.E.A - Digital Edition of Archilochus' fragments</publisher> <pubPlace>Parma</pubPlace> <date>2018</date> <availability> <p>This fragment is available only for demonstration purposes. (user license)</p> </availability> </publicationStmb>
Source Description	<sourceDesc> <bibl>
Title	<title type="Volume">Archilochus, Hipponax, Theognidea</title>
Author and Fragmet	<title type="Part">Archilochus, fr. 23 W.<chi rend="apex">2</chi></title>
Meter	<note>Iambic Trimeters</note>
Ancient Author	<author xml:id="Archilochus">Archilochus</author>

Figure 2

TEXT	
Section	<div type="tipo_di_sezione_es:titolo" cert="000"></div>
Number of the verse	<xml:Id="L1">
Gap (uncertain length)	<gap reason="lost" extent="unknown" unit="chars"/>
Gap (certain length)	<gap reason="lost" quantity="00" unit="chars" cert="grado_di_certezza"/>
Lines lost	<gap reason="lost" extent="unknown" unit="lines"/>
Letters (illegible)	<gap reason="illegible" quantity="2" unit="chars"/>
Reading (unclear)	<unclear>α</unclear>
Supplement (uncertain letters)	<supplied reason="illegible">ι</supplied>
Supplement (lost letters)	<supplied reason="lost">ι</supplied>
Text deleted (ancient witness)	<del rend="erasure">αβ
Text deleted and illegible (ancient witness)	<del rend="erasure"><gap reason="lost" quantity="3" unit="character"/>
Text added (ancient witness)	<add place="000">αβ</add>
Text deleted (modern editor)	<surplus>αβ</surplus>
Text added (modern editor)	<add resp="#editore">μév</add>

Figure 3

In this model, that can be applied to both manuscript texts and papyrus (or epigraphy) texts, the metadata include general information as bibliographical sources, manuscript or papyrus (or epigraphy) description, available open data (as imagines) and, in the <body>, we encode gap, lacunae, different reading, corrections, symbol and then apparatus with scholars' reading and hypothesis.

Finally, we can use the results obtained in this small but complex field of study to create a replicable model for other fields of literary studies. We can also identify standards and best practices for the enrichment of our digital edition with structured knowledge from the Semantic Web. The investigation of technological aspects should hopefully act as an important test for future projects in the field of classical studies.

2.2 CLARIN-IT and DEA project: data and metadata

Digitization of Archilochus' texts allows the creation of a philologically and critically controlled product. The project is aimed to integrate the available digital resources, implementing and enriching what already exists and to develop crucial resources, materials and tools for study and research. The project also represents one of the attempts within the context of CLARIN-IT⁵ to contribute to the wider impact of CLARIN (Common Language Resources and Technology Infrastructure) on the specific Italian community interested to DC and may improve services in fostering new (and sustaining existing) knowledge in SSH digital research.⁶

As Beatrice Nava (Nava, 2019) points out in her Tour de CLARIN Interview: "DEA can be regarded as a case study in the framework of CLARIN-IT and its interests and specialization towards the Digital Classics. [ILC4CLARIN](https://ilc4clarin.ilc.cnr.it/) (https://ilc4clarin.ilc.cnr.it/) offers the corpus in their repository, along with other existing digitized resources for Ancient Greek (e.g. a Linked Open Data - LOD)⁷. This allows us to enrich our corpus with lexical datasets in LOD and to integrate our data with other existing resources, with the final aim of obtaining a complete edition that are useful not only for scholars interested in Classical and Ancient Studies but also for non-specialist users."

Moreover, as she said: "linguistic annotation allows the development of new teaching methods of Ancient Greek that are aimed at beginners and include the use of language services such as treebanks and tools like [TüNDRA](https://weblicht.sfs.uni-tuebingen.de/Tundra/) (https://weblicht.sfs.uni-tuebingen.de/Tundra/), but specifically implemented for classics. Perhaps more generally, the annotations enable an interactive approach to texts that is more inviting and immediate to the students. [...] In our case, having a closed corpus allows us to develop linguistic services for teaching (e.g., [Hyper-Text Archilochus](#))⁸." To sum up, improving the currently existing parsers for Ancient Greek with regards to their performance on fragmentary texts would offer very important upgrades for the study and teaching of Ancient Greek.

At this stage we test some cases (encoding text and metadata) and we store them in CLARIN-IT. We think that these texts will be important for the design and construction of the CLARIN infrastructure because we can find metadata and concept registries, and we can catalogue and browse

⁵ See (Monachini and Frontini 2016).

⁶ See (Monachini, Nicolosi and al. 2018).

⁷ Version of the TEI-dict Perseus Liddell-Scott Jones Greek-English dictionary (<http://lari-datasets.ilc.cnr.it/ml/>).

⁸ A prototype that provides the learner with a set of resources and tools that ease a critical assessment of ancient texts (<http://hdl.handle.net/20.500.11752/OPEN-83>).

data. Moreover, we use a Clarin Mobility Grant to advance our studies and our knowledge⁹; it was a clear example of how CLARIN can develop (and hopefully improve) researcher training activities¹⁰. In addition, we can consider that an experience like that of Tour de CLARIN can spread Infrastructure Knowledge and can be fundamental for the Dissemination of research projects.

3 Conclusion: addressing a specific research challenge

Research infrastructures seems to be the perfect place to make the results obtained concrete and visible. This project provides an example of how it is possible to integrate and support the proof-reading, encoding and enrichment of the texts in a repository. This newly created resource will be integrated into the existing exploration system and, in order to be researchable, all texts will be provided with adequate metadata.

Finally, we can make ours the wishes of Beatrice in her Interview. It would be helpful that CLARIN-IT introduce in its repositories an integrated online environment that would support the proof-reading, encoding and enrichment of classical texts. This can improve CLARIN-IT with guidelines to provide metadata specific to digital classics. Moreover, would be useful that CLARIN-IT develop tools tailored to non-computational researchers that would help them perform linguistic and textual annotation (morpho-syntactic, semantic, etc) without requiring them to possess a great deal of technical know-how.

References

- Carmen Brando, Francesca Frontini, and Jean-Gabriel Ganascia. 2016. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *Complex Systems Informatics and Modeling Quarterly*, 0 (7): 60–80.
- Monica Monachini and Francesca Frontini. 2016. CLARIN, l'infrastruttura Europea Delle Risorse Linguistiche per Le Scienze Umane e Sociali e Il Suo Network Italiano CLARIN-IT. *IJCoL - Italian Journal of Computational Linguistics, Special Issue on NLP and Digital Humanities*, 2 (2): 11–30.
- Monica Monachini, Anas Fahad Khan, Francesca Frontini and Anika Nicolosi. 2018. Linked Open Data and the Enrichment of Digital Editions: The Contribution of CLARIN to the Digital Classics. In *Proceedings of the CLARIN Annual Conference 2018*, edited by Inguna Skadina and Maria Eskevich. Pisa, Italy. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf.
- Monica Monachini, Anika Nicolosi and al. 2018. Digital Classics and CLARIN-IT: What Italian Scholars of Ancient Greek Expect from Digital Resources and Technology. In *Selected Papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, 61–74. Linköping University Electronic Press, Linköpings universitet.
- Beatrice Nava. 2019. *Tour de CLARIN: Interview with*, <https://www.clarin.eu/blog/tour-de-clarin-interview-beatrice-nava>
- Anika Nicolosi. 2013. *Archiloco: elegie*. Pàtron Editore, Bologna, Italy (ISBN 9788855532365). Google-Books-ID: 9uj5oAEACAAJ.
- Anika Nicolosi. 2015. *Analisi testuale e linguistica di Lirici Arcaici e Adespoti Giambici ed Elegiaci: Ipotesi di ricerca di applicazione della Filologia Computazionale al Greco Antico*. ILC-CNR, Pisa, Italy, November 6. http://www.ilc.cnr.it/sites/default/files/presentations/ILC-Thematic-Seminar_11.06.2015_Presentation.pdf
- Anika Nicolosi. 2017. *Archiloco. Testimonianze e frammenti*. Aracne Editrice, Roma, Italy (ISBN 9788825508550).

⁹ See CLARIN blog (<https://www.clarin.eu/blog/tei-and-ancient-greek-fragmentary-poetry>).

¹⁰ Seminar was given also in Parma (<https://www.clarin.eu/blog/clarin-it-presents-their-roadshow-seminars>).

Research Data of a PhD Thesis Project in the CLARIN-D Infrastructure. “Texts of the First Women’s Movement” / “Texte der ersten Frauenbewegung (TdeF)” as Part of the German Text Archive

Anna Pfundt
Justus Liebig University
Giessen, Germany
Anna.Pfundt@germanistik
.uni-giessen.de

Melanie Grunt Suárez
Justus Liebig University
Giessen / University of
Tübingen, Germany
Melanie.Grunt-
Suarez@germanistik.
uni-giessen.de

Thomas Gloning
Justus Liebig University
Giessen, Germany
Thomas.Gloning@germanistik
.uni-giessen.de

Abstract

The authors of this paper are going to present parts of a PhD thesis, that examines the use of words in the German discussion on the controversy of women’s suffrage around 1900. The study refers to a variety of written texts (including journal articles, books, and controversial writings) that began to condense in the 1880s and developed a complex thematic network until the introduction of women’s suffrage in 1918. The focus of this paper is the presentation of the corpus compilation (ongoing and already published to some extent) for the CLARIN-D infrastructure component German Text Archive (“Deutsches Textarchiv”, hereafter DTA). This project addresses a basic user need, to make new texts available from the very beginning of a project. Each new text increases the material basis for the dissertation, which can be analysed with the powerful search tool architecture of the DTA. On the other hand, the textual repertoire of the DTA grows with each text. Finally, it’s a win-win situation both for the author, for the infrastructure and for the whole research community.

1 Historical Text Corpora in the CLARIN Infrastructure: A Use Case from a Dissertation Project on the Word Usage in the Controversy over Women’s Suffrage 1870-1918

The subject of Anna Pfundt’s PhD thesis project is the use of words in the discussion about women’s suffrage from the 1870s until 1918. This debate was conducted in a broad spectrum of texts (including journal articles, books, and controversial writings) that began to condense in the 1880s and developed a complex thematic network until the introduction of women’s suffrage in 1918. The use of words plays a central role in the constitution of points of view, in the formulation of views and their justification, as has often been the case in discourses about alternative and competitive word usages (see Pfundt, 2017; Gloning, 2012). The use of words is characterised by its own thematic profile, but also by the controversial nature of the object and by aspects of historical development over several decades. In addition to the specific uses (“meanings”) of words and phrases, forms of word formation, metaphors, ad hoc uses, foreign-language expressions and functionally oriented vocabulary sectors are also part of the study. Another important aspect for the project is the connection of the use of words to different “camps” and their perspectives and opinions.

The research is methodologically based on the work of the Düsseldorf School around Dietrich Busse, Georg Stötzel and Martin Wengeler (e.g. “Kontroverse Begriffe”, ed. by Stötzel and Wengeler, 1995). It is characterised by a discourse historiography in which the thematic developments of a discourse are reconstructed in narrative form in connection with the description of the lexical means used. In the investigation, the narrative presentation of the discussion and the use of words anchored in it is based on a corpus of German texts on early women’s suffrage, the texts of which are successively fed into the DTA.¹ By way of illustration we present a short example from an early text by Hedwig Dohm, one of the earliest champions of suffrage in Germany:

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

O über dieses Geschwätz von der **Sphäre** des Weibes, den Millionen Frauen gegenüber, die auf Feld und Wiese, in Fabriken, auf den Straßen und in Bergwerken, hinter Ladentischen und in Bureaus im Schweiß ihres Angesichts ihr Brot erwerben. Wenn die Männer vom weiblichen Geschlecht sprechen, so haben sie dabei nur eine ganz bestimmte Klasse von Frauen im Sinn: Die Dame. Wie nach dem bekannten Ausspruch jenes bekannten österreichischen Edelmannes der Mensch erst bei dem Baron anfängt, so fängt bei den Männern das weibliche Geschlecht erst da an, wo es Toilette und Conversation macht und Hang zu Liebesintrigen und Theaterlogen verräth. Geht auf die Felder und in die Fabriken und predigt eure **Sphärentheorie** den Weibern, die die Mistgabel führen und denen, deren Rücken sich gekrümmt hat unter der Wucht centnerschwerer Lasten! (Dohm, 1876, p. 126-127)

In this passage, Hedwig Dohm criticizes the assumption of different spheres of men and women, an assumption which was common in 19th century thinking and which nowadays is called “difference assumption” (Differenzannahme, Differenzhypothese).

As a third component, the dissertation contains an extensive discourse dictionary, which currently contains almost 1100 entries. The three components (investigation, text corpus, discourse dictionary) and their intertwining is shown schematically in the following figure (Wolff et al., 2015, figure 6):

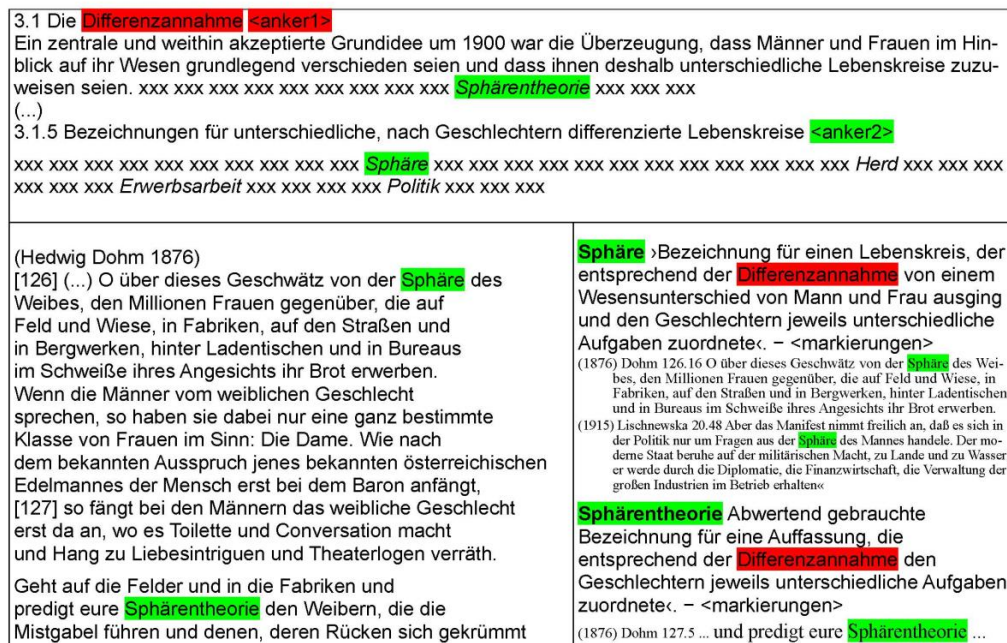


Figure 1: Connections between investigation, text corpus and discourse dictionary.²

With regard to the text corpus component, a win-win situation arises for the author of the dissertation and for the CLARIN-D infrastructure component DTA. Each new text increases the material base, which

¹<http://deutschestextarchiv.de/> (last access 2019-08-23).

²The figure is not a suggestion for screen design but rather a visualization of the structural connections between investigation, dictionary and corpus texts.

can be analysed with the powerful DDC³ search tools of the DTA. For the DTA, each of these texts is a thematic enrichment and – with regard to texts written by women – also a contribution to increasing the proportion of female authors.

2 Texts from the First Women's Movement around 1900 in CLARIN's German Text Archive

The empirical investigation is based on the DTA sub corpus "German Texts of the First Women's Movement" / "Texte der ersten Frauenbewegung" (hereafter TdeF)⁴. It collects German-language texts from the realm of the women's movement around 1900, at present mainly with a thematic reference to women's suffrage. In total 64 texts and 2159 pages are already published.⁵ This is certainly not yet a huge corpus, but a valuable contribution of texts written by women and dealing with highly important topics for the emancipation of women.

The corpus is limited to the period 1870-1919. Since the first texts on the issue of women's suffrage appeared around 1870 and women were granted the right to vote in 1918, the historical period is clearly defined. Since the central media at that time were books, magazines and newspapers, the corpus texts are independent works in the form of books, brochures or lectures as well as magazine articles from the press organs of the women's movement. The authors are both opponents and supporters of the women's movement, they are organisations, men and women, some anonymous.

In order to create the texts as searchable full texts in a corpus and make them publicly accessible, they are digitised according to the specifications of the DTA basic format.⁶ The DTA basis format has become a quasi-standard for the encoding of texts from the 17th century onwards. For lexical analysis, the texts in the subcorpus TdeF can be addressed as a separate text group for queries. The texts are produced either with ABBYY Recognition Server which is able to deal with blackletter typefaces, or are keyed in manually according to the quality and complexity of the images. Still in other cases transcriptions on the web can be used as a starting point. In all cases meticulous proofreading is necessary.

The discourse dictionary on the one hand serves to analyse and to document individual word usages, on the other hand the entries may be addressed from passages in the investigation in order to present further details and textual quotations. The entries are compiled to provide a pragmatic-semantic description of the individual lexical units and their specific uses. The dictionary serves as a lexicographical documentation of the lexical inventory of the discourse. The terms that are important in the history of this discourse and that are commented on in the thematic word usage profiles of the study are also found in the dictionary. The work on this digital discourse dictionary is based essentially on the TdeF corpus and is to be linked directly to the full texts via the references of the documents. TEI guidelines are applied for the encoding of the digital dictionary as well. This dictionary is supposed to be published in the ZHistLex project⁷ by one of the CLARIN partners as well.

3 The Contribution of the CLARIN Corpus-Infrastructure for a Prototypical Use Case: Text Corpus, Investigation, Discourse Dictionary

The TdeF corpus is the basis for the empirical study; it provides the textual basis for both the discourse dictionary and the investigation of word usage. Since the full texts can be searched and analysed using the DDC search engine of the DTA, the search for words and phrases as well as the overview of frequency relationships and word-formation correlations is considerably facilitated.

As a prototypical use case, the dissertation project shows how a lexical investigation can simultaneously benefit from the offers of the CLARIN-D infrastructure (here the DTA) and contribute to its further expansion. The compilation of the TdeF corpus is at the same time the successively expanded textual basis for the dissertation and an important contribution to expand the holdings of the DTA in an important phase of German language and discourse history.

³http://deutschestextarchiv.de/doku/DDC-suche_hilfe

⁴<http://www.deutschestextarchiv.de/doku/textquellen#tdef>

⁵<http://www.deutschestextarchiv.de/dtae>

⁶<http://www.deutschestextarchiv.de/doku/basisformat/>

⁷<https://zhistlex.de/>

4 What we Will Show in Leipzig

In Leipzig, we will present this use case with a poster, focusing in particular on the interaction of the infrastructure component (DTA) with the scientific goals of the dissertation project. In view of the win-win situation mentioned above, this use case could be a model for other dissertation projects or lexicological investigations of a similar nature. It can have an advertising effect and invite people to share their data from the beginnings of a project in a similar way. Comparable specialised corpora are in preparation (e.g. early film documents; historical cookery recipes; early herbals).

References

Hedwig Dohm. 1876. *Der Frauen Natur und Recht* [About the nature and the rights of women]. Wedekind & Schwieger, Berlin. http://deutschestextarchiv.de/book/show/dohm_frauenfrage_1876 (last access 2019-08-25).

Thomas Gloning. 2012. Diskursive Praktiken, Textorganisation und Wortgebrauch im Umkreis der ersten Frauenbewegung um 1900. *Historische Pragmatik*. Ed. Peter Ernst. De Gruyter, Berlin, New York. 127-146.

Anna Pfundt. 2017: Frauenwahlrecht? Oder Damenwahlrecht? Oder doch ein allgemeines Wahlrecht? – Zum Wortgebrauch in der Diskussion um das Frauenwahlrecht um 1900. *Im Zentrum Sprache*, 2 November 2017: <https://sprache.hypotheses.org/542>. Last access 2019-08-23.

Georg Stötzel and Martin Wengeler. 1995. *Kontroverse Begriffe. Geschichte des öffentlichen Sprachgebrauchs in der Bundesrepublik Deutschland*. De Gruyter, Berlin, New York.

Kerstin Wolff, Alexander Geyken, and Thomas Gloning. 2015. Kontroverse Kommunikation im Umkreis der ersten Frauenbewegung. Wie können digitale Ressourcen die sprachliche Untersuchung und die Ergebnisdokumentation verbessern? *Grenzen und Möglichkeiten der Digital Humanities*. Eds. Constanze Baum and Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). text/html Format. http://zfdg.de/sb001_010. Last access 2019-08-23.

Appendix: The Profile Description of the TdeF Corpus (Transl. from German)

<http://www.deutschestextarchiv.de/doku/textquellen#tdef>

The corpus “Texts of the First Women’s Movement” / “Texte der ersten Frauenbewegung (TdeF)” consists of German-language source texts that originated in the women’s movement around 1900. The issue of women’s suffrage is currently in the foreground, but in the controversial discussion that has taken place between the various camps of the women’s movement, individual representatives and the parties in parliament, many other contemporary issues have also been raised. Since the first texts on the issue of women’s suffrage appeared around 1870 and women were granted the right to vote in 1918, the historical period is clearly defined. The texts are independent works in the form of booklets or brochures as well as journal articles published in the press organs of the women’s movement. The collection is supervised by Anna Pfundt and Thomas Gloning, Justus Liebig University Giessen. The full texts are captured according to XML/TEI P5 according to the DTA basic format and published in the DTA.

Granularity versus Dispersion in the Dutch Diachronical Database of Lexical Frequencies TICCLAT

Martin Reynaert	Patrick Bos / Janneke van der Zwaan
DHLab & Meertens Institute	Netherlands eScience Center
KNAW Humanities Cluster, Amsterdam	Amsterdam, The Netherlands
& Tilburg University – The Netherlands	p.bos@esciencecenter.nl
reynaert@uvt.nl	j.vanderzwaan@esciencecenter.nl

Abstract

The Nederlab project collected the digitized diachronical corpora of Dutch and made them available to all researchers in a single, explorable and exploitable portal within the CLARIN infrastructure. We are now building a database of lexical items and their frequencies collected according to the best known year of text production or publication on the basis of the 18.5 billion word tokens in the corpus. We here briefly discuss the corpus contents, major database design decisions we have taken, the tools we use and the approaches we take.

1 Introduction

We¹ have worked in ‘spelling correction’, very broadly put, for going on for two decades. The great paradox we see in non-words in texts, whether having been created by mistyping or misrecognition by some text digitization system or any other mishap, is that when they have been resolved to the real-word that ‘should be there’, they have in the large majority of cases been solved, once and for all. This was in fact the vision behind the work on spelling correction already performed by IBM researchers in the 1980s (Pollock and Zamora, 1984), who advocated ‘absolute correction’, i.e. if a known error is encountered: replace it by its correct form. In Reynaert (2005) we gave an example of the non-word ‘onjections’ that might variously have to be resolved to ‘injections’, perhaps given the context ‘these painful *onjections’, versus to ‘objections’ given the context ‘her vehement *onjections’. At least for longer words, measured in numbers of characters, such ambiguities are in fact rather rare.

The above gives the main rationale behind the current project TICCLAT, which stands for ‘Text-Induced Corpus Correction and Lexical Assessment Tool’. It is meant to help assess the validity of word forms encountered in Dutch diachronical text. Its databases, or selected subsets thereof, will assist OCR post-correction, but a great many more uses may easily be envisaged. We want to use the vocabulary present in what is today the largest finely preprocessed corpus of Dutch, actually a compilation of many corpora, collected in the prior project Nederlab (Brugman et al., 2016), to try and solve most of these non-words, once and for all. To resolve them and link them to their most likely real-word versions, then to have this database available online² to all comers, freely usable for whatever research purposes the community may find uses for. Apart from this, we hope to greatly enhance the historical lexica available to us with as many as possible of the historical spelling variants ever produced as reproduced in these subcorpora. Thirdly, we want to account for the morphological variants of words, starting with contemporary Dutch, gradually going back over time to at least the 13th. century.

In fact, the way we are proceeding is to start with the best contemporary resources we have, to augment these with the evidence we encounter in the (re-)born-digital corpora we have and lastly to proceed to the much larger, but far and far noisier OCR-digitized corpora we have in the Nederlab corpus. The current estimate is that only around 6% of the 18.5 billion word tokens of text in Nederlab is born-digital.

¹Here: the first author.

²The TICCLAT database is to be hosted by Clarin Center Meertens Institute (<https://centres.clarin.eu/centre/23>)

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

2 From corpora to year-stamped frequency lists

Corpora Overview: the Nederlab corpora Table 1 gives an overview of the main Nederlab subcorpora ingested in TICCLAT.

Period	Corpus Title	Type	Size
13th. century	“Walewein ende Keye”	Book	S
13th. to 21st. cent.	DBNL: Digital Library of Dutch Literature	Books	L
14th. century	Corpus of 14th. Century Dutch by Van Reenen & Mulder	Acts	M
17th. century	Scholarly Correspondences (Geleerdenbrieven and Epistolarium)	Letters	M
1620-1640	Minutes of the States of Holland (by N. Stellingwerff and S. Schot)	Reports	M
1618 to 1700	KB or Dutch National Library Newspaper Collection (crowd-sourced by Nicoline van der Sijs, rekeyed)	Newspapers	L
1701 to 1940	KB or Dutch National Library Newspaper Collection (OCR)	Newspapers	H
1621-1700	Acta of the Particular Synodes of Southern Holland	Reports	M
1693-1701	Pieter van Dam’s “A description of the Dutch East India Company (VOC)”	Book	M
17th. and 18th. cent.	“Prize Papers” aka: “Sailing Letters” (edited by N. van der Sijs)	Letters	M
1702-1720	Correspondences of Anthonie Heinsius	Letters	S
1780-1800	EDBO: Early Dutch Books Online	Books	L
1811-1831	Diaries of Willem de Clercq	Diaries	S
1814 - 2014	Dutch Acts of Parliament (Political Mashup version)	Reports	H
1874-1918	Diaries and Notes of Willem Hendrik de Beaufort	Diaries	S
19th. century	Vincent van Gogh – The Letters	Letters	S
1891-1947	Diaries of P.J.M. Aalberse	Diaries	S
1985-2005	STEVIN Written Dutch Reference Corpus SoNaR-500	30 genres	L

Table 1: Overview of the main subcorpora available in Nederlab. Size estimates: S = Small, M = Medium, L = Large, H = Huge.

Method We want to account as best possible for as many of the lexical items as automatically as possible in terms of real-word versus non-word, contemporary real-word versus diachronical variant and name versus non-name. We first account for morphologically related word forms. These should give us a first handle on true word forms likely to be expected in a language, regardless of the time frame the particular language is inspected at. TICCL will next be used in line with new developments achieved since Reynaert (2011) for identifying historical variants and to account for non-word variants.

Granularity and Dispersion The main issue here is granularity. Our means to zoom in on the data is obviously to obtain frequency lists containing the lexical items and their frequencies. But, given e.g. the EDBO, should we get the overall corpus frequency (‘how often does this word form occur in EDBO?’), the document frequency (‘in how many EDBO books does this word form occur?’), its frequency for each book of the corpus, or even as the Hathi Trust provides for books in its collections, its frequency per page of each book in the corpus? What we currently opt for is yet another take: per subcorpus, as far as the metadata allows, we regard the documents originating in the same year (or range of years in case the exact year of text production or publication is not known) as belonging to each year’s subsubcorpus and collect this subsubcorpus frequency for all its word types.

Having the corpus frequencies for the range of all the years of all subcorpora, we are then enabled to get a clear and humanly interpretable overview of the occurrence of a particular word form across time as well as across the corpora that make up the full Nederlab corpus. In this manner, the first link we have established between word forms in the TICCLAT database is that of their dispersion (Baayen, 1996) over time and over a range of diverse corpora. The term dispersion, “i.e. the degree to which occurrences of a word are distributed throughout a corpus evenly or unevenly/clumpily” is further qualified by Stefan Th. Gries in a new book chapter³ as “one of the most crucial but at the same time underused basic statistical measures in corpus linguistics”. Note that we have added the notion of subcorpora of the ‘corpus’, further subdivided into year-stamped further subsubcorpora, and that we wish to study and compare or contrast the distribution of ‘words’ over these. The main challenges we face are the result of the highly uneven quality of the text collections we work with, resulting in inordinate numbers of (non-)word types.

³Preprint at: https://www.researchgate.net/publication/332120488_Analyzing_dispersion

Impact of available metadata In this work we are at the mercy of the quality of the data and metadata available for each text and we encounter difficulties concerning this in e.g. the DBNL, which should by rights be considered one of the most valuable digital text collections extant for Dutch. The DBNL was largely built before (Moretti and Piazza, 2005) introduced the notion of Distant Reading. No usable distinction was made between e.g. contemporary commentaries and quoted historical text fragments. Also, the metadata available to us for at least this subcorpus is unsatisfactory to our purposes. Mediaeval works republished in recent years that have no mention of a text production year anywhere near the lifetime of the original author will definitely impact the reliability of our time-stamped word frequencies. Metadata is likely to remain problematic. Whoever compiles a corpus cares to record to the best of her abilities the best available metadata, given the goals of her project. Given the likely very divergent goals of later projects, given the availability of the particular corpus to the larger research community with different research interests, the available corpus metadata is unfortunately all too likely to be found wanting. We carry on regardless.

Means: the word frequency tools used We strive to get the most usable overview over the diachronical lexicon of Dutch as present in a wide and diverse range of word lists and digital text collections. These corpora are available to us in FoLiA XML and we have the tools to highly efficiently and in the best parallelized fashion process these (van Gompel et al., 2017). There are millions of files, however. So what we did was not bring the data to the tools, but rather bring the tools closer to the data. This is not yet what currently the Dutch CLARIAH-plus project works towards, i.e. infrastructure to bring the tools to the data, which can then remain in its proper repository and does not need to be copied and distributed, possibly leading to multiple copies differently (pre-)processed, transformed, linguistically enriched, etc. What we have now is a means of virtually reordering the otherwise stationary file collections for further processing, e.g. on the basis of metadata such as ‘date of publication’ or ‘location’ or any other criterion. We have extended two of the C++ TICCL modules⁴ in order to extract the frequency lists from the Nederlab subcorpora in the best way possible.

FoLiA-stats derives frequency lists from corpora. FoLiA-stats – there is also a TICCL-stats which works on plain text – in its original working mode can recursively traverse directory trees, locating the files to be jointly transformed into word ngram frequency lists. To do what we want to, this implies we would have had to copy all the files with text originating in the same year to a single directory for one year, which on its own and separately from the other years, would have to be processed by FoLiA-stats. We brought the tool closer to the data by extending it so that it can work on the basis of a list containing the full directory paths and file names of the thousands, for some subcorpora: millions, of texts to be processed. A label in the second column instructs the program to create a directory with the same name and to collect the vocabulary contained in all the files bearing the same label in a single frequency file to be output to that directory. So, regardless of the actual whereabouts of the files regarding directories or even storage partitions, a single frequency list containing the ngrams observed in the texts that originated in a single year can be created, in parallel, for a range of years.

The second tool, **TICCL-unk**, is next enlisted to ‘clean’ the word types from the corpus frequency list compiled by FoLiA-stats. Especially in OCRed corpora, which are untokenized, character strings occur that are, heuristically, deemed unsalvageable, i.e. OCR garbage. These are disregarded. Character strings having ‘word’ initial or final punctuation are written to another file and shorn of this punctuation added to the main list where, if already present, their corpus frequency is added to that of the clean version. Clean word strings are naturally written to the ‘clean’ file and their frequencies tallied.

3 The TICCLAT database

Structure of the TICCLAT database and tools The structure of the TICCLAT database is squarely adopted from the historical lexical database structure our project partner INT⁵ developed in the European project IMPACT⁶. We have extended it with a number of fields required for our purposes, including

⁴Available from: <https://github.com/LanguageMachines>

⁵Institute for the Dutch Language: <https://ivdnt.org/the-dutch-language-institute>

⁶<https://www.digitisation.eu/>

being able to specify links between wordforms. An overview of the database schema and link to the IMPACT document can be found in the documentation⁷. The **TICCLAT software**⁸ is used for database management, and ingesting and querying the data.

On linking based on word forms' relatedness It is one thing to fill a huge database with hundreds of year-stamped frequency lists of subcorpora derived from both born-digital and OCR-digitized texts. It is quite another to organize these many millions of word forms in a sensible way to start seeing through the trees and make them useful, whatever the intended use. In TICCLAT, we link the many related variants of what might constitute a single 'word'. Each word type is assigned a unique code which links and identifies through its prefix the overall cluster of its related words, by an infix specific to each of the following three subcategories: the morphologically related word forms, next the word types related diachronically or that are possibly divergent but accepted word variants and, finally, the incredible diversity of related erratic word forms misrecognized by the digitization processes. Numerical suffixes identify the word clusters and each unique word form in the cluster. The supervised morphology induction system we have developed to derive these codes in itself warrants a full paper to discuss the related work and provide a full evaluation.

4 Conclusions

What we have sketched is in fact a huge undertaking. Were it not that we actually have control over the granularity of both setting the timeline of the subsubcorpora we ingested in the database and that of the subcorpora we choose to TICCL in order to extract the lexical variants from and best-first rank these by, we might not achieve anything noteworthy in the limited time frame of the TICCLAT project. We are setting up the infrastructure for demonstrating on a sufficiently large scale that the ultimate goal, given the necessary resources, is achievable.

Acknowledgements

The authors acknowledge being funded in the call 'Accelerating Scientific Discovery in the Arts and Humanities' (ADAH) issued jointly by CLARIAH and the Netherlands eScience Center.

References

- Harald Baayen. 1996. The effects of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics*, 22:455–480, 12.
- Hennie Brugman, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Antal van den Bosch. 2016. Nederlab: Towards a Single Portal and Research Environment for Diachronic Dutch Text Corpora. In Nicoletta Calzolari et al., editor, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC-2016)*, pages 1277–1281, Portoroz, Slovenia. ELRA.
- F. Moretti and A. Piazza. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.
- Joseph J. Pollock and Antonio Zamora. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of the ACM*, 27(4):358–368.
- Martin Reynaert. 2005. *Text-Induced Spelling Correction*. Ph.D. thesis, Tilburg University.
- Martin Reynaert. 2011. Character confusion versus focus word-based correction of spelling and OCR variants in corpora. *International Journal on Document Analysis and Recognition (IJ DAR)*, 14(2):173–187.
- Maarten van Gompel, Ko van der Sloot, Martin Reynaert, and Antal van den Bosch. 2017. FoLiA in practice: The infrastructure of a linguistic annotation format. In J. Odiijk and A. van Hessen, editors, *CLARIN-NL in the Low Countries*, chapter 6, pages 71–81. Ubiquity (Open Access).

⁷https://github.com/TICCLAT/docs/blob/master/database_design.md

⁸<https://github.com/TICCLAT/ticclat>

Cross Disciplinary Overtures with Interview Data: Integrating Digital Practices and Tools in the Scholarly Workflow

Stefania Scagliola

University of Luxembourg
stefania.scagliola@uni.lu

Louise Corti

University of Essex
corti@essex.ac.uk

Silvia Calamai

University of Siena
silvia.calamai@unisi.it

Norah Karrouche

Erasmus University Rotterdam
karrouche@eshcc.eur.nl

Jeannine Beeken

University of Essex
jeannine.beeken@essex.ac.uk

Arjan van Hessen

University of Twente
a.j.vanhessen@utwente.nl

Christopher Daxler

University of Muenchen
daxler@phonetik.uni-muenchen.de

Henk van den Heuvel

Radboud University
H.vandenHeuvel@let.ru.nl

Max Broekhuizen

Erasmus University Rotterdam
maksbroekhuizen@gmail.com

Abstract

The progress in computer science with regard to capturing and interpreting forms of textual human expression is impressive. This does however not seem to be the case when considering the standard scholarly approach to interview data: speech of two or more persons that is turned into text to be studied and analysed. To set the stage for assessing the potential integration of new technology in this field, a community of experts from the Netherlands, Great Britain, Italy and Germany who engage with interview data from different perspectives, organized a series of workshops that were funded by CLARIN (Oxford, Utrecht, Arezzo, München, and Utrecht; 2016-2019). This paper sketches the organization of the workshops, the preliminary results and envisioned further lines of research. It presents the goals and the selection of participants, data and tools, reflects on how the invited scholars coped with unfamiliar approaches and digital tools and describes how in the next stages efforts will be made to include new languages, new open source annotation tools, and conducted research behavior. A multilingual archive of oral history interviews covering the topic of migration brought together by the organizers, was the basis for the first exploration, and will be used for further experiments to assess whether and how cross-disciplinary collaboration and the exchange of methods, data and tools can lead to innovation in methodology, use and services.

1 Developing a Transcription Chain

The first goal of the transnational project was to develop a portal for automatic transcription and alignment of interview data for different languages. The participating countries and languages were selected on the basis of the availability of mature open source speech retrieval software: German, English, Dutch and Italian. The development of such a portal was taken up by the group of speech technologists and phonologists within the research network. Participants were recruited that represented the following communities: historians and social science scholars who undertake research with recorded interview data; linguists who use spoken language resources and information scientists and data archive curators who are responsible for access to and curation of interview data. During the first three workshops transcription practices and user requirements were documented on the basis of the performance of various speech to text software on interview data. This had been provided by researchers and data curators in the four different languages. Clarifying the different workflows of

scholars was a key requirement for the development of a Transcription-chain (Speech-to-Text software and audio and text alignment) that could cater for multiple needs. For further information on the building of the T-Chain the reader is referred to van den Heuvel et al (2019) and to the portal: <https://oralhistory.eu/workshops>.

2 Cross disciplinary overtures in München (2018) and Utrecht (2019)

The fourth workshop, held in München (19-21 September 2018) created the opportunity to fully test the prototype of the T-Chain and its interface. For this purpose each language group had provided audio data in their own language enabling to test the T-Chain from a multilingual point of view. This time however, the goal was to go beyond the phase of transcription, and consider technologies that could be used in the subsequent phases of the research process: annotation and analysis. These sessions were attended by a balanced mix of British, Dutch, German and Italian scholars, who all used different tools on the same audio corpus: a collection of interviews from the 1980s, on black migrants from the Caribbean to Great Britain. Anticipating that the diversity of participants and tools would make the organization of the workshop complex, it was essential to follow principles in the design of the workflow that ensured ‘satisfying user and research experiences’:

1. gathering Information on the participants level on their level of digital shrewdness
2. prepare data familiar to the participants in both a common language (English) and in their native language
3. assign homework in order to make participants become familiar with the tools,
4. ensure that a participant with advanced technical skills was present in each of the language groups
5. collect feedback directly after each workshop session exercises, in order to improve the approach a next time

In the first session the T-Chain was tested using German, Dutch, English and Italian data. Automatic speech recognition being an essential component in the chain, it was stressed that such engines work no miracles, but can be helpful in obtaining a usable transcription in a more efficient way than transcribing from scratch, if the material is of good recording quality, and preferably offered in segments of, say, 15 minutes. Also audio-data conversion and segment selection via the tools *GoldWave* and *Audacity* were demonstrated. In the subsequent three sessions participants worked with proprietary and open source annotation tools, common among social scientists (*ELAN* and *NVivo* software), with OS and commercial linguistic text mining tools used by computational linguists (*VOYANT*, *Stanford CoreNLP*, *TXM* software), and with emotion recognition tools used by computer scientists/linguists (*PRAAT*, *OpenSmile* software).

Before starting the hands-on sessions it was deemed necessary to present the research profiles of the disciplines represented in order to counter the risk of ‘simplification’ that occurs when referring to other disciplines. It is almost trivial to observe that within every discipline distinct sub-disciplines exist. Speaking of ‘linguistics’ is an over-simplification, also the term ‘oral history’ covers a broad variety of approaches. An oral historian will typically approach a recorded interview as an intersubjective account of a past experience, whereas a colleague might consider the same source as a factual testimony. A social scientist is likely to try to discover common themes and similarities and differences across a whole set of interviews, whereas a computational linguist will do the same, but based on counting frequencies, detecting collocations and co-occurrences, keywords in context (KWIC), positive and negative scoring, and constructing (conceptual) metaphorical/semantic networks. To illustrate the variety of landscapes, participants were invited to provide ‘research trajectories’ that reflected their own approach(es) when working with interview data. This enabled us to come up with a high-level simplified trajectory and to identify how and where the digital tools might fit into the researchers’ workflow.

2.1 Tools for transcription, annotation, linguistic analysis and emotion recognition

Researchers were invited to work in four ‘language groups’ of 5 to 6 people (Dutch, English, Italian and German) in hands-on sessions, using step-by-step worksheets and pre-prepared interview extracts. The T-Chain, developed with CLARIN support, with its speech to text and alignment software, was able to

partly substitute the cumbersome transcription of interviews, a practice that is common to anyone working with interviews. When trying out the annotation tools, that offer a structured system to attribute meaning to passages, the common needs tended to decrease. The reason is that a choice of a particular annotation tool leads to an engrained practice of research that cannot be easily traded for an alternative. For this purpose the open source tool *ELAN* was used, and compared with the proprietary qualitative data software *NVivo*.

In the following two sessions the experimental component increased, as the same interview data was explored with computational linguistic tools of increasing complexity. First the web-based (OS) tools *Voyant* and *NLPCore* were used allowing the processing of transcripts ‘on the fly’, thus enabling a whole set of specific language features to be directly identified. The second tool, *TXM*, had to be installed and allowed for a more granular analysis of language features, requiring the integration of a specific language model, the splitting of speakers, the conversion of data into computer readable XML language, and the lemmatization of the data. The last session was the most daunting one, illustrating what the same data yielded when processed with the acoustic tool *PRAAT*, and the facial recognition tool *OpenSmile*.

2.2 User experience and evaluation

The goal of the workshop was to give all participants an overview of what is available and useful, outside one’s own research domain, and to open their minds to the advantages of different technologies. Many scholars were not familiar with the method and terminology (e.g. text mining), let alone the tools. A first analysis of the user experience before, during and after the workshop suggests that scholars are open to cross-fertilization. At the same time, scholars are only willing to integrate a digital tool into their existing research practice and methodological mindset, if it can easily be used or adapted to their needs. The limited functionality of the free easy-to-use tools, and the observed methodological and technological complexity and jargon-laden nature of the dedicated downloadable tools, were both seen as significant barriers, despite the availability of clear documentation.

It was clear when assessing the annotation tools *ELAN* and *NVivo*, that researchers struggled with the unfamiliar interface and lingo. When probing what exactly the obstacles were, it appeared that *NVivo*, as a tool that has been designed for a broad spectrum of uses, despite the clear instructions, did not connect to the specificity of the participants research practice. A whole lot of ‘translation work’ was needed, in terms of skill development and training, in order to really successfully affect the mindset of the participants. The broad potential of *NVivo* was recognized, but the tool was perceived as complicated, or “too powerful”, as one historian remarked. *ELAN*, a free annotation tool that contrary to *NVivo* was specifically designed for linguistics, led to confusion, as it did not make a clear distinction between functionalities to create a transcription or to make annotations.

The text mining tools faced an additional challenge, as the added value of sifting out patterns or the use of grammar and lexicon, together with word frequency are not at all obvious outside the realm of linguistics. When presented with more simple tools, such as the web-based text mining tool *Voyant* or *CoreNLP* and different autosummarizers, specific use cases were described by the researchers. Historians did see potential in broadly exploring text data as a first step before close-reading. However, the simplicity of these tools (just cut a piece of text, paste it in a window and press on ‘process’) triggered questions about the limited ‘transparency’ of the tool, of not knowing what exactly happens ‘under the hood’.

What was harder to deal with, was the linguistic tool *TXM* that required data to be preprocessed. Such tools have a high learning curve. It appeared difficult to understand how to attribute meaning to the frequency of a particular term in the entire corpus of interviews, when considering a single interview. The same applied to the emotion recognition tools. The messy data that preprocessing interviews about migration yielded, led to sifting out possible hypotheses, but not to a deeper understanding of the experience of migration. The real challenge lies in being able to translate insights with regard to scale, frequency and paralinguistic features into the classic interpretation of the interview data. Often this means looking at other features, for instance the amount and nature of silences and emotion within an entire corpus (Truong et al 2013; Van den Heuvel & Oostdijk, 2016).

The most salient conclusion of the exploration could be that the traditional practice of interpretation of

interviews could be enriched by considering digital tools as purely heuristic instruments, used to explore the data at the very beginning of the research process, i.e. when one is still considering what collections to reuse or what approach to take. Another key conclusion was how jargon hampers the uptake of technology. Explanation of linguistic approaches would be better appreciated in more lay terms, following a step-by-step pathway from meta-level to concrete level. A third concern was related to the ethical issues of uploading an interview in a web service without knowing what happens to this data. This prompted thinking about how these tools could add explicit GDPR-compliant data processing agreements to allay worries about whether or not the data remains on the servers after being processed by e.g. the speech recognizer. From this respect, a closer collaboration with the CLARIN Ethical and Legal Committee is envisioned. Finally, users pointed to a great need for contextual information about, and metadata for, the data collection and processing, when interview data sources are used. Inviting language resources and scholars across languages and different disciplines certainly enriched the meeting experience.

3. Conclusion

The paper suggests that interdisciplinary or multidisciplinary conversations about interview data help to reveal new, innovative uses of digital tools and new methods for and approaches to analysis. But there is still a lot of work to do in the realm of ‘translation’. Accordingly, the transnational research group aims at pushing forward the Technology and Tools for Oral History initiative. Therefore, some of the envisioned next steps are:

- a further elaboration of the T-Chain with more languages;
- a further uptake of open source annotation tools;
- a research study in which a computational linguist and a historian study/analyse the same corpus and compare/evaluate their results;
- a research study in which we process all the evaluations we collected, that provides a useful snapshot of how scholars can better communicate/innovate across disciplines

The multilingual archive of oral history on migration created by the research team will be the basis for further explorations, in order to assess in which way technology and protocols for data collection (metadata schemes, tagging, specific topic-lists) can make oral history archives useful for a variety of scholars; to verify whether and how cross-disciplinary collaboration, exchange of methods, data and tools, can lead to innovation in methodology; to assess how automatic transcription services can be best structured to meet the variety of requirements of scholars that work with oral history data; to investigate whether in the near future also the communities of archivists, librarians and curators should/can be involved in the research.

References

- Khiet Truong, Gerben J. Westerhof, Sanne M.A. Lamers, Franciska de Jong, and Anneke Sools. 2013. Emotional expression in oral history narratives: comparing results of automated verbal and nonverbal analyses. *Proceedings of the Workshop on Computational Models of Narrative CMN 2013 Hamburg, Germany*.
- Henk Van den Heuvel, Christopher Draxler, Arjan Van Hessen, Louise Corti, Stef Scagliola, Silvia Calamai, and Norah Karouche. 2019. A Transcription Portal for Oral History Research and Beyond. *Proceedings DH2019, Utrecht, 10-12 July 2019*.
- Henk Van den Heuvel, and Nelleke Oostdijk. 2016. Falling silent, lost for words ... Tracing personal involvement in interviews with Dutch war veterans. *Proceedings LREC2016, Portorož, Slovenia: 998-1001* http://www.lrec-conf.org/proceedings/lrec2016/pdf/104_Paper.pdf

The Rise of the Definiteness Effect in Icelandic

Einar Freyr Sigurðsson

The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
einar.freyr.sigurdsson@arnastofnun.is

Anton Karl Ingason

University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

This paper looks at the Definiteness Effect (DE) in the history of Icelandic and argues, using the Icelandic Parsed Historical Corpus (IcePaHC), that DE in its current form is relatively recent. This is in line with Ingason et al. (2013) who argued that DE played a crucial role in the development of the so-called New Impersonal Passive in Icelandic.

1 Introduction

The Definiteness Effect (DE) has been argued to play an important role in the development of the New Impersonal Passive (NIP) in Icelandic (Eythórsson, 2008; Ingason et al., 2013). The DE applies in the Canonical Passive (CanP), see (1), whereas it does not in the NIP, see (2). That is, what makes (1) ungrammatical is the fact that the definite noun phrase (NP) stays in situ whereas the accusative case NP in (2) can be definite without affecting the grammaticality of the sentence.¹

- (1) **Það var lamið stúlkan.*
EXPL was beaten.F.NOM the.girl.F.NOM
Intended: ‘The girl was beaten.’ (Eythórsson, 2008, 177)

- (2) %*Það var lamið stúlkuna í klessu.*
EXPL was beaten.DFLT the.girl.F.ACC in a.mess
‘The girl was badly beaten.’ (Maling and Sigurjónsdóttir, 2002, 98)

Eythórsson (2008) suggested, discussing the emergence of the NIP, that a “leakage” in the DE led to reanalysis of the CanP with a definite postverbal NP. Furthermore, Indriðadóttir (2014), interpreting results of her own questionnaire, proposed that the DE is on the decrease in Modern Icelandic. Ingason et al. (2013), on the other hand, argued that the **rise** of the DE was a crucial factor in the spread of the NIP. We take this to be an unresolved issue. The purpose of the current paper is to examine quantitative facts about the DE in the history of Icelandic, using the Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. 2011), to better evaluate different accounts of the evolution of the DE and the NIP.

2 The Definiteness Effect and the NIP

2.1 DE in Modern Icelandic

As discussed by Milsark (1977), English existential constructions are subject to a definiteness restriction.

- (3) There is **a wolf** at the door.
(4) *There is **the wolf** at the door.

This restriction, standardly referred to as the Definiteness Effect (DE), applies to Icelandic as well, as shown in (5)–(6). In the grammatical example (5), the indefinite NP *úlfur* ‘a wolf’ does not move to the subject position. If the NP is definite, as in (6), the sentence is ungrammatical.²

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

¹The NIP is rather widespread among younger speakers even though it is strictly ungrammatical for many speakers, especially among the older generations. Therefore we use the ‘%’ sign to indicate that only some speakers accept the NIP.

²It should be noted that the expletive *það* in Icelandic is a first-position element and not a thematic subject.

- (5) *Það er úlfur við dyrnar.*
EXPL is a.wolf.NOM at the.door
'There is a wolf at the door.'
- (6) **Það er úlfurinn við dyrnar.*
EXPL is the.wolf.NOM at the.door
Intended 'The wolf is at the door.'

The DE applies in various constructions, such as the existential construction in (5)–(6) as well as the Canonical Passive (CanP), see (7)–(8).

- (7) *Það var lesin bók.*
EXPL was read.F.NOM a.book.F.NOM
'A book was read.'
- (8) **Það var lesin bókin.*
EXPL was read.F.NOM the.book.F.NOM
Intended: 'The book was read.'

For (6) and (8) to be grammatical, the definite NP must move to subject position, as shown in (9)–(10).

- (9) *Úlfurinn er við dyrnar.*
the.wolf.NOM is at the.door
'The wolf is at the door.'
- (10) *Bókin var lesin.*
the.book.F.NOM was read.F.NOM
'The book was read.'

However, DE does not apply in the New Impersonal Passive, as we will now see.

2.2 The lack of DE in the NIP

The New Impersonal Passive (NIP) is a construction with passive morphology NP without movement to subject position, whether or not the NP is definite. By comparing the NIP in (11) and the CanP in (8) and (10), we can see that the DE does not apply in the NIP but only the CanP (see, e.g., Maling and Sigurjónsdóttir, 2002). This suggests that the status of the theme argument is different in the NIP than in the CanP. It should also be mentioned that the NP is assigned objective case in the NIP, accusative in (11).³ This differs from the CanP, see (10), where a NP that is assigned accusative case in the active is in the nominative case in the passive.

- (11) *%Það var lesið bókina.*
EXPL was read.DFLT the.book.ACC
'The book was read.'

Note that the lack of DE is not a general feature of NIP speakers in other constructions than the NIP (Maling and Sigurjónsdóttir, 2002). However, Indriðadóttir (2014) tested a few constructions with an expected DE violation. 92 adolescents in 10th grade were tested and the results were compared to the responses of 15 speakers at the age 65–75 years. Indriðadóttir comes to the conclusion that the restrictions set by the DE have weakened for some speakers and that it is weaker for the younger speakers than the older speakers. There may be an ongoing change in this direction in Modern Icelandic but it is nonetheless not consistent with the development of the DE in the history of Icelandic, it seems, as discussed in §3.

3 The emergence of the DE

3.1 DE as a factor in the emergence of the NIP

For an account of the emergence of the NIP, Eythórsson (2008) looked at cases where the CanP and the NIP are the same on the surface. He suggests that there may be exceptional DE violations in CanP input in language acquisition, such as in (12) where *litla barnið* is syncretic for nominative and accusative; nominative would reflect a DE violation in the CanP whereas accusative would reflect the NIP.

- (12) *%Það var skammað litla barnið.*
EXPL was scolded little the.child.NOM/ACC
'The little child was scolded.'
(Eythórsson, 2008, 181)

That is, examples like (12) may occasionally be generated by a CanP grammar, even though it should not be possible given DE. Eythórsson refers to this as a "leakage" in the DE. Attested examples of such

³Maling and Sigurjónsdóttir (2002) argued that the NIP contains a projected implicit argument; see also H.Á. Sigurðsson (2011), Ingason et al. (2013), Legate (2014), E.F. Sigurðsson (2017).

leakage are found at various times in the history of Icelandic. Eythórsson (2008, 183) shows the following example of DE leakage with a postverbal definite NP in the dative case from the 13th century *Sturlunga saga* (note that accusative case in structures with passive morphology, as in (2) and (11) above, always reflects the NIP but dative and genitive can be ambiguous between CanP and NIP grammars).

- (13) *Var ýtt skipinu.*
 was pushed the.ship.DAT
 ‘The ship was pushed.’

(Eythórsson, 2008, 183; *Sturlunga saga*)

The crucial leakage has to do with dative and genitive case, rather than nominative case, as only the former (dative, genitive) can be (re-)interpreted as being produced by an NIP grammar.

However, if DE leakage was a necessary factor for the rise of the NIP, the question is why the change did not take place in, say, Old Icelandic. We could expect DE leakage to be more frequent in the 20th century, when the change caught on, than in older stages of Icelandic. Ingason et al. (2013, 98), on the other hand, argue that a leakage in the DE was not the important factor in the emergence of the NIP:

“The NP owes its advantage to the definiteness effect. Therefore, it was the rise of the definiteness effect, not its leakage, that created favorable conditions for the spread of the New Passive. While this does not explain why the first NP speaker acquired the new grammar, it does predict that such an innovation had no chance of spreading before the 20th century.”

Furthermore, Ingason et al. (2013, 97) state that “there was no categorical definiteness effect until the 20th century”. This could mean that the DE was not a part of the grammar of speakers or, alternatively, that a grammar with the DE constraint was applied with a probability lower than 100% at the relevant stages of the history of the Icelandic language, i.e., the DE was non-categorical at the time.

3.2 DE in the history of Icelandic

To evaluate whether there has been any change in the history of the Definiteness Effect in Icelandic, we look at quantitative data from the Icelandic Parsed Historical Corpus (IcePaHC), which contains around 1 million words from the 12th century through the 21st century.

We focus on definite NPs following passive participles and take a look at relative frequencies over the time period covered by IcePaHC. These are shown in Figures 1–2. In our search queries, we looked at definite NPs, annotated as subjects, in the dative and genitive case, on the one hand, and in the nominative case, on the other, following a passive participle out of all passives with definite dative and genitive case subjects and definite nominative subjects, respectively (for more information on querying IcePaHC, see Ingason, 2016). It seems clear that the relative frequency is higher as we go further back in the history of Icelandic. Therefore, it looks like Ingason et al. (2013) are right when they say that there was a rise in the Definiteness Effect (with decreasing relative frequency of apparent DE violations over time). The figures suggest that in earlier Icelandic there may not have been such a phenomenon as DE.

If a DE leakage is relatively infrequent in Modern Icelandic, as Figures 1–2 suggest, it is not clear what kind of circumstances are needed for re-interpretation. We suggest that at a certain point in history, the evidence that children are exposed to during language acquisition with respect to the DE does not warrant exceptions (or a leakage) anymore and thus a categorical DE emerges in the language.

4 Conclusion

Data from IcePaHC support Ingason et al.’s (2013) account regarding the Definiteness Effect and the New Impersonal Passive rather than the leakage hypothesis as proposed in Eythórsson (2008). Furthermore, the results suggest that there was not an active DE rule in earlier periods of Icelandic. We argue that DE leakage is too frequent in Old Icelandic such that there was no actual DE at the time. It is, however, difficult to interpret the results with respect to individual grammars: We do not know why the postverbal definites decreased steadily. We also do not know whether some speakers at earlier periods had a DE in their grammar while others did not. We leave these speculations for future research which should also take into account the change from OV to VO word order and control for weight effects.

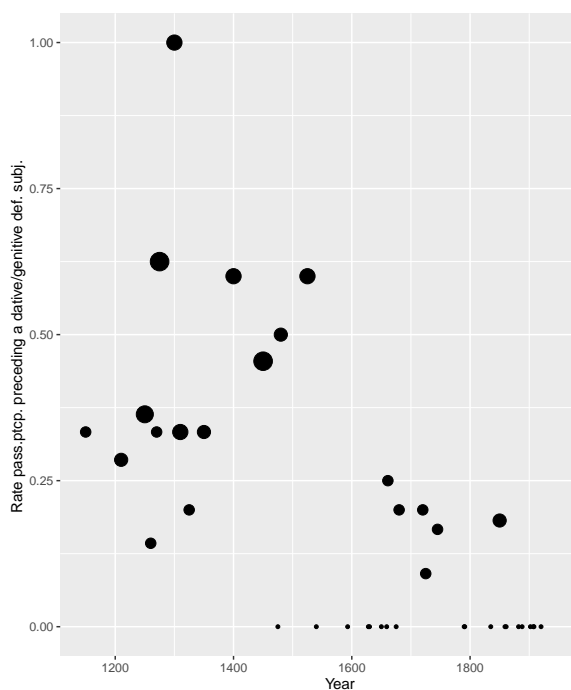


Figure 1: Relative frequency of dative and genitive definite subjects following a passive participle (out of all passives with def. dative and genitive subjects).

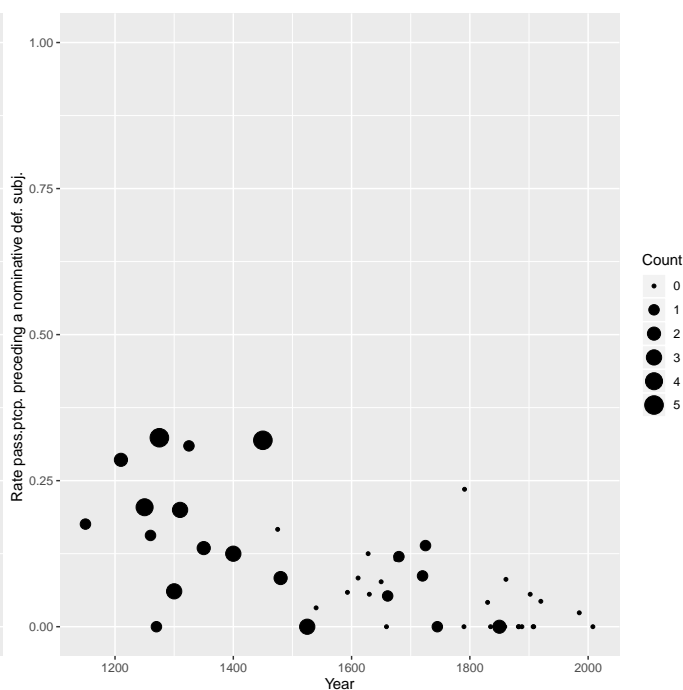


Figure 2: Relative frequency of nominative definite subjects following a passive participle (out of all passives with definite nominative subjects).

References

- Eythórssón, Thórhallur. 2008. The New Passive in Icelandic really is a passive. Thórhallur Eythórssón (ed.): *Grammatical Change and Linguistic Theory. The Rosendal papers*, pp. 173–219, John Benjamins, Amsterdam.
- Indriðadóttir, Ingunn Hreinberg. 2014. *Er búin mjólkinn? Hamla ákveðins nafnliðar og tengsl hennar við nýju setningagerðina*. M.A.-thesis, University of Iceland, Reykjavík. <http://hdl.handle.net/1946/17762>
- Ingason, Anton Karl. 2016. PaCQL: A new type of treebank search for the digital humanities. *Italian Journal of Computational Linguistics* 2(2):51–66.
- Ingason, Anton Karl, Julie Anne Legate and Charles Yang. The Evolutionary Trajectory of the Icelandic New Passive. *University of Pennsylvania Working Papers in Linguistics* 19(2):91–100.
- Legate, Julie Anne. 2014. *Voice and v: Lessons from Acehnese*. MIT Press, Cambridge, MA.
- Maling, Joan, and Sigríður Sigurjónsdóttir. 2002. The ‘new impersonal’ construction in Icelandic. *Journal of Comparative Germanic Linguistics* 5:97–142.
- Milsark, Gary. 1977. Toward an explanation of certain peculiarities of the existential construction in English. *Linguistic Analysis* 3:1–29.
- Sigurðsson, Einar Freyr. *Deriving case, agreement and Voice phenomena in syntax*. Doctoral Dissertation, UPenn.
- Sigurðsson, Halldór Ármann. 2011. On the New Passive. *Syntax* 14:148–178.
- Thráinsson, Höskuldur. 2007. *The Syntax of Icelandic*. Cambridge University Press, Cambridge.
- Wallenberg, Joel C., Anton Karl Ingason, Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson. 2011. Icelandic Parsed Historical Corpus. Version 0.9. http://www.linguist.is/icelandic_treebank

Integrated Language and Knowledge Resources for CLaDA-BG

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

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

Abstract

This paper presents the envisaged integration of the language resources for Bulgarian with the knowledge sources like ontologies and linked open data to support their joint usage with respect to the cultural and historical heritage (CHH) objects. We started with the knowledge integration of the language resources for Bulgarian. Our plan is to continue with the addition of selected CHH objects to the initially integrated data. Based on the available Bulgarian resources like dictionaries and corpora as well as on the Bulgarian Wikipedia, DBpedia and Wikidata, we have constructed the first version of a Bulgaria-Centric Knowledge Graph. It represents the conceptual information for the Bulgarian virtual infrastructure CLaDA-BG.

1 Introduction

Nowadays vast networks with linked objects are dominant in all areas of life, including tools and data in NLP. Among the many prominent initiatives in linking available data in various combinations are the following: the Linked Open Data Cloud¹ (language resources with ontologies), the Predicate Matrix² (a lexical resource that integrates the information from different semantic and syntactic resources such as FrameNet, VerbNet, PropBank, WordNet), PARTHENOS project³ (integrates cloud storage with services and tools, and supports collaborative working on language and CHH data), SSHOC⁴ (connecting existing and new infrastructures from the SSH ERICs).

CLaDA-BG is the Bulgarian National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies.⁵ In contrast to other EU infrastructures that started separately as CLARIN and DARIAH, and later on in some countries (Austria, the Netherlands) combined or started to work in closer cooperation, in Bulgaria the joint infrastructure started as such from the beginning. In the spirit of European CLARIN and DARIAH, the mission of CLaDA-BG is to establish a national technological infrastructure of language resources and technologies (LRT), and cultural and historical heritage (CHH) resources and technologies. The consortium of CLaDA-BG comprises 15 organizations including research institutes at the Bulgarian Academy of Sciences, several universities, the National Library “Ivan Vazov” in Plovdiv, and two museums. Thus the consortium includes not only technological partners, but also content providers and some of the users of CLaDA-BG as a research infrastructure.

The main goal of the infrastructure is to provide public access to these resources and technologies for various societal tasks, targeted at a wider audience. The infrastructure aims to support

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

¹ <https://lod-cloud.net/>

² <http://adimen.si.ehu.es/web/PredicateMatrix>

³ <http://www.parthenos-project.eu/>

⁴ <https://sshopencloud.eu/>

⁵ <http://clada-bg.eu/>

predominantly researchers in Art, Humanities and Social Sciences to process Bulgarian language texts and CHH datasets necessary for their research. In order to support this type of research, we would like to put the varying types of data in the context of each other. The approach for interlinking of the data is called *contextualization*. The different types of objects of study, representation and search are integrated on the basis of common metadata categories and via textual descriptions. The language resources and the textual descriptions of other objects are integrated with the help of a common Bulgaria-Centric knowledge graph - *BGKG*.

Here we present the core set of language resources that in our view is necessary to support the research in SSH. Their integration is discussed with respect to the semantic annotation of texts with conceptual information from the knowledge graph. The ultimate goals are: the extraction of new knowledge from text, querying the knowledge graph, and indexing of texts in the CLaDA-BG repository.

2 Integrated Bulgarian Language Resources

Within the CLaDA-BG plan for Language Resources and Technology we follow the notion of Basic Language and Resource Kit – BLaRK – (Krauwert 2003). Here is the initial list of the envisaged language resources. The language processing tools will include minimally the following ones: morphological, shallow syntactic, deep syntactic, and semantic analyzers, named entities recognition and identification modules. The basic language resources are:⁶

- Text Archive for Bulgarian (minimum 100 million running words), searchable on Internet;
- Morphologically Annotated Corpus (1 million running words);
- Syntactically annotated corpus (1 million running words);
- Semantically annotated corpus with ontological and fact information (1 million running words);
- Bulgarian Wordnet (BTB-WN) (50 000 synsets of coverage of the lemma senses in a related semantically annotated corpus)
- Valency lexicon (coverage of the verbs in BTB-WN)
- Inflexional lexicon – currently 110 000 lemmas with their paradigms
- Domain corpora and related dictionaries (minimum 100 000 running words per domain)
- Representative lists of Bulgarian names (coverage of the names of the public figures, location and organization names. Additionally, they will include relations to the Bulgarian Wikipedia)

During the first year of the CLaDA-BG project we focused on the integration of the various existing language resources and performed only minimal extension in order to make them usable. As a basis for the manually annotated corpus we use the texts included in BulTreeBank – an HPSG-based treebank for Bulgarian – comprising about 260 000 running tokens. These texts were annotated before the start of CLaDA-BG with senses from BTB-WN and instances from DBpedia, URLs from Wikipedia and classes from DBpedia ontology (see Popov et al., 2014). In Figure 1 we present an example of a sentence from BulTreeBank annotated with senses and named entities from BTB-WN and Bulgarian DBpedia.

The sentence is: “*Водещ на купона беше Тома Спространов.*” (“The host of the party was Toma Sprostranov.”) The two open class words are connected with the respective synsets from BTB-WN, represented here by their definitions. The word “*Водещ*” (“The host”) is a present participle of the verb “*водя*” (“to organize”) and it is annotated with that sense of the verb. From the fact that it is a participle, present tense, it follows that the word denotes the person who organizes the event. The word “*купона*” (“the party”) is connected to the definition “*Организирано увеселение ...*” (“Organized entertainment”). The host was the disco jockey Toma Sprostranov who has a Wikipedia page: https://bg.wikipedia.org/wiki/Тома_Спространов

In the cases when there is no Wikipedia page for the corresponding named entity, we add only a class from the DBpedia ontology, such as Person, Politician, Musician, Country, City, Document, etc. Note that beside the original constituent-based annotation the dependency representation is supported.

⁶ In brackets we put the desirable minimal size of the corresponding resource.

The dependency annotation follows the Universal Dependency guidelines.⁷ The original Treebank is also manually annotated on morphosyntactic level with a rich tagset (680 tags – Simov et al., 2004) and lemmas. Thus, the extended annotation of the treebank with senses, named entities from Wikipedia, and conceptual information from DBpedia provides data for training and testing of modules for Word Sense Disambiguation, Named Entities Recognition and identification.

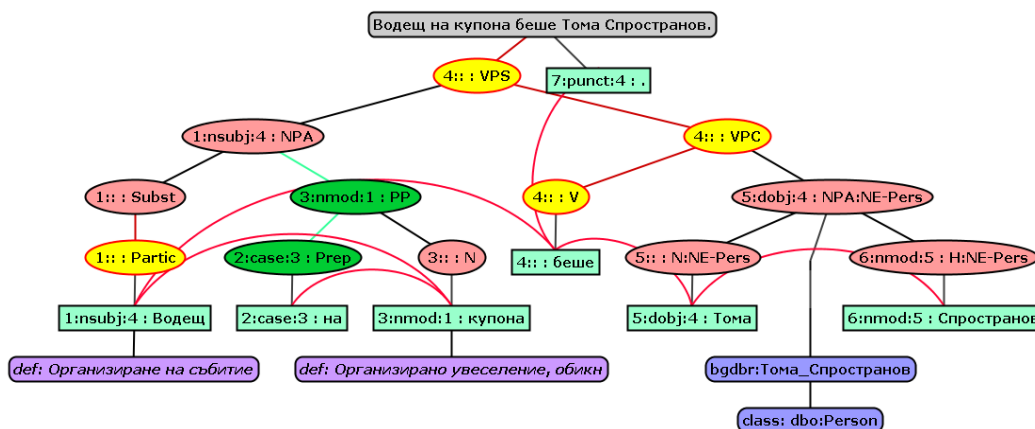


Figure 1: Example of a BulTreeBank sentence annotated with senses from BTB-WN (word forms 1 and 3) and a named entity linked to Wikipedia/DBpedia (word forms 5 and 6)

The main lexical resources within the CLaDA-BG language resource toolkit (Bulgarian WordNet BTB-WN, Valency lexicon, and Inflectional lexicon) are also in a process of integration. The BTB-WN contains 22 000 synsets which cover all the words within BulTreeBank and most frequent words over the Bulgarian national reference corpus - BulTreeBank (about 100 million running words). We started with the extension of the information within BTB-WN by adding inflectional paradigms to each lemma in the synsets and with their mapping to articles within Bulgarian Wikipedia to synsets. The inflectional information is important because many lemmas in BTB-WN belong to different inflectional paradigms. Thus even the current quite precise lemmatizer cannot ensure a correct word sense assignment for some word forms. The verbs in the Valency lexicon are mapped to the corresponding synsets from BTB-WN. Each verb is associated with valency frames depending on its meaning. Each frame element is also mapped to the synsets in BTB-WN.

The information from Bulgarian Wikipedia provides not only additional encyclopedic information for the named entities, but also terminological one. During the process of mapping of BTB-WN to the Wikipedia, new senses have been entered to the lexical resource. In addition, the mapping to Wikipedia provides a source for new relations between the synsets in BTB-WN – see Simov et al., 2019. On the basis of the current 22 000 synsets in BTB-WN, we extracted a little more than 13 000 Wikipedia articles. These articles were manually inspected and mappings between the synsets and the articles were established. The mapping follows the approach of McCrae 2018. In addition to the Wikipedia articles that correspond to the synsets in BTB-WN, we selected and extracted 10 899 Wikipedia articles that relate to the names in a Bulgarian gazetteer. This gazetteer represents the most important names in the Bulgarian National Reference Corpus. From them 1 515 pages were already extracted on the basis of the lemmas within BTB-WN. The rest 9 384 pages were classified as Bulgarian locations, other locations, people, organizations, and other. In this way we extend BTB-WN with important for Bulgaria named entities.

The integration of the language resources will be used at least in two directions (1) training of a wider set of processing modules, and (2) contextualization through the relations from the text to the encyclopedic information. The latter is considered very important for the connection between language processing and suitable information extraction from textual descriptions of cultural and historical objects.

⁷ <https://universaldependencies.org/>

The integration of language resources and encyclopedic knowledge is a first step into the direction of constructing a knowledge graph for CLaDA-BG aligned to the language resources for Bulgarian.

3 Towards a Bulgaria-Centric Knowledge Graph

We aim at creating a semantically integrated environment for maintaining possibilities of referring to texts and descriptions of cultural or historic objects. For this purpose the texts and descriptions of collections should be annotated with an appropriate ontology and then the annotation will be uploaded into an RDF repository. The first version of BGKG is based on Bulgarian DBpedia knowledge graph and in the process of implementation of CLaDA-BG we will add knowledge from other sources. Besides Wikipedia and DBpedia we envisage the inclusion of Wikidata as part of the initial knowledge graph. The integration of these sources of knowledge is ensured by their design. Wikidata as a knowledge source is with a higher level of quality ensured by rigorous rules for construction and manual inspection.

The integration of the initial knowledge graph with cultural and historic objects is a non-trivial task, since it will have to ensure dynamic flow of information from and to databases like Europeana, Wikidata and local to the provider classifications which are often not formalized.

In future a unified nomenclature dictionary will be developed for managing the various types of data: texts, images, artefacts, but for the first version of the knowledge graph we will start with the integration of cultural and historical objects provided by the partners within CLaDA-BG. The integration with Europeana and Wikidata will be tested through several pilots. The description files will be checked and corrected as well the relations and digital objects. The metadata will be enriched with information about the organization and the aggregator. Example queries will be generated, such as: maps of selected integrated objects, picture tiles, bubble charts over various topics, etc. In addition to the manual mapping we will process their textual descriptions which will provide the connection between the knowledge graph, the metadata for the described objects and the mentions of the corresponding instances and classes in the text. For example, from texts we could extract information like the material, technique, etc. used in the creation of a given artefact.

The knowledge graph will be available via a search tool. The search tool will provide the following search possibilities: a) concept search; b) facet search (integrating several concepts) and c) combined search (integrating concepts with random key words). This will ensure similar search for mentions of conceptual information in the texts and in the semantic description of the cultural and historical objects. The inclusion of the language, cultural and historical information into a common knowledge graph will provide one of the main mechanisms to support the research through the contextualization of each object of analysis.

4 Conclusion

In the abstract we present the ongoing development of language and semantic resources within CLaDA-BG to support research in Humanities and Social Studies. This is done through the exploration of text corpora and the description of cultural and historical objects. In the area of language resources, the integration of Bulgarian language data through BTB-WN and the Bulgarian Wikipedia provides a basis for training of text indexing with instances and classes from a Knowledge Graph. The actual knowledge graph is based on DBpedia and Wikidata (including also information from Wikipedia). Besides the textual information, descriptions of cultural and historical heritage objects are expected to be mapped to the knowledge graph as well. This step will allow a joint search including a SPARQL endpoint. When developed enough, the knowledge graph will be provided freely for download as a linked open dataset. The initial knowledge graph will be further extended by specific ontologies for modelling the specific classification schemata, or time and space, events, facts.

Acknowledgements

This research was partially funded 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 DO01-164/28.08.2018.

References

- Krauwier, Steven. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In Proceedings of the 2nd International Conference on Speech and Computer (SPECOM2003).
- McCrae J. P. 2018. Mapping WordNet Instances to Wikipedia. In Proceedings of the 9th Global WordNet Conference (GWC 2018), pp. 62–69.
- Popov, A., Kancheva, S., Manova, S., Radev, I., Simov, K., Osenova, P. 2014. The Sense Annotation of BulTreeBank. Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), 2014.
- Simov, K., Osenova, P., Slavcheva, M. BTB-TR03: BulTreeBank Morphosyntactic Tagset. BulTreeBank Project Technical Report № 03. 2004.
- Simov, K., Osenova, P., Laskova, L., Radev, I., and Kancheva, Z. 2019. Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. The 10th Global WordNet Conference. Wroclaw, Poland.

Application of a Topic Model Visualisation Tool to a Second Language

Maria Skeppstedt^{1,2,*}, Magnus Ahlthorp¹, Andreas Kerren³, Rafal Rzepka^{2,4}, Kenji Araki²

¹The Language Council of Sweden, the Institute for Language and Folklore, Sweden
{maria.skeppstedt,magnus.ahlthorp}@sprakochfolkminnen.se

²Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
{rzepka,araki}@ist.hokudai.ac.jp

³Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
andreas.kerren@lnu.se

⁴RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Abstract

We explored adaptations required for applying a topic modelling tool to a language that is very different from the one for which the tool was originally developed. The tool, which enables text analysis on the output of topic modelling, was developed for English, and we here applied it on Japanese texts. As white space is not used for indicating word boundaries in Japanese, the texts had to be pre-tokenised and white space inserted to indicate a token segmentation, before the texts could be imported into the tool. The tool was also extended by the addition of word translations and phonetic readings to support users who are second-language speakers of Japanese.

1 Introduction and background

Topic modelling provides a means of extracting a relevant subset of documents from collections that are too large to make a fully manual analysis of all its documents feasible. The extracted documents are organised into groups by the topic modelling algorithm, each group corresponding to a topic that occurs frequently in the document collection. This ability to, in an unsupervised fashion, extract and topically sort relevant documents has been used to perform qualitative text analysis in social science and humanities research (Baumer et al., 2017).

We have previously presented a tool for visualising the output of topic modelling, which we call Topics2Themes (Skeppstedt et al., 2018a; Skeppstedt et al., 2018b). There are several tools for visualising topic modelling output, for instance with the focus on assessing and improving the quality of the topic model produced (Chuang et al., 2012; Lee et al., 2012; Jaegul Choo et al., 2013; Hoque and Carenini, 2015; Lee et al., 2017; Smith et al., 2018; Cai et al., 2018), and with the focus on supporting the user in exploring and interpreting the texts included in the document collection (Alexander et al., 2014).

Although the output of topic models in the form of a selection and sorting of documents has been shown useful for speeding up and facilitating qualitative text analysis, previous research has shown the need for users to identify other pieces of information than the automatically created topics. For instance, to identify reoccurring themes in the texts extracted, which are more thematically detailed than automatically identified topics (Baumer et al., 2017). Topics2Themes, therefore, does not only include functionality for letting the user explore and interpret the automatically extracted topics and documents, but places an equal emphasis on allowing the user to add, and subsequently explore, an additional layer of analysis. This is carried out by enabling the creation of user-defined themes that can be associated with the documents extracted by the topic modelling algorithm.

We originally created Topics2Themes for English texts. Despite the unsupervised nature of the topic modelling algorithm, which makes Topics2Themes fairly language-independent, it is not self-evident that the tool can be applied as-is to text written in a language that is typologically very different from English. To investigate this, we applied the tool to texts written in Japanese, i.e., a language that is both morphologically and orthographically different from English.

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

* International Research Fellow of Japan Society for the Promotion of Science (Postdoctoral Fellowships for Research in Japan (Short-term))

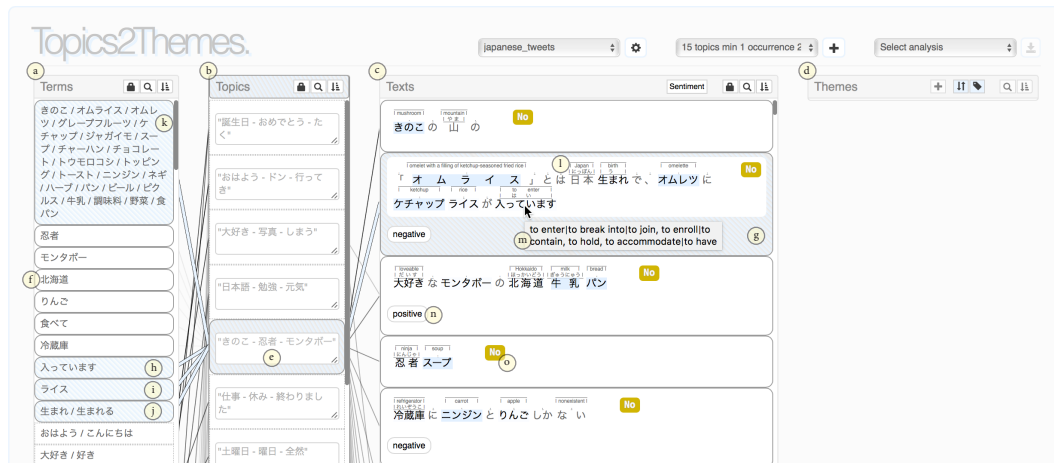


Figure 1: (a-d) The *Terms/Topics/Texts/Themes* panels. (e) The selected topic. (f) Example of rounded border indicating terms and texts associated with the selected topic. (g) The text over which the mouse hovers. (h-k) Terms associated with the text over which the mouse hovers. (k) Cluster of food-related words. (l) Language support in the form of phonetic reading and English translation. (m) Additional English translations for the word over which the mouse hovers. (n) A static label attached to the text by the word list matching. (o) A label that the user can change, here given an initial neutral value.

In addition, we envisioned the situation in which the text analysis of the Japanese texts would be performed by an analyst that would require some level of language support for fully understanding the texts. Such a situation would most naturally occur in a language learning situation, i.e., a situation in which the interaction with the documents is the primary reason to use the tool, and the output of the analysis is only of secondary importance. This situation could, however, also occur in the case in which a second-language speaker needs an understanding of the important content of a document collection, without having the means of employing the help of a more proficient speaker of the language. With this situation in mind, we incorporated a system into Topics2Themes that helps second-language speakers of Japanese to understand Japanese text.

2 Adaption to Japanese and the addition of reading support

Topics2Themes¹ uses a very simple tokenisation based on the occurrence of white space. As white space is not normally used in Japanese to indicate word boundaries, another technique for tokenisation is required. We decided not to change the tokenisation method built into Topics2Themes, but to instead require the texts imported into the tool to be pre-tokenised and white space inserted into the texts to indicate token segmentation. The tokenisation included in Topics2Themes could therefore be used as-is. For this pre-processing, we applied segmentation using the MeCab tool (Kudo, 2006), and then merged segments to tokens by matching them to the JMDict dictionary (JMDict, 2013), as implemented by Ahltop (2012).

We also configured the tool to use Japanese stop words², instead of using English ones, as well as to use a word2vec model trained on a Japanese corpus to perform concept clustering. That is, Topics2Themes can be configured to apply dbSCAN clustering on word2vec vectors corresponding to the words in the corpus, and let all words belonging to the same cluster be collapsed into one concept, before the text is submitted to the topic modelling algorithm. For performing the clustering on Japanese, we configured Topics2Themes to use vectors from a word2vec model³ that had been trained on Japanese texts, which

¹The code for the Topics2Themes tool is available at: <https://github.com/mariask2/topics2themes>, and the code for the Japanese pre-processing can be obtained by contacting the authors.

²We used stopwords from: <https://github.com/stopwords/japanese-stopwords/blob/master/data/japanese-stopwords.txt> and extended them by frequent non-content words in the corpus used.

³<https://github.com/shiroyagicorp/japanese-word2vec-model-builder>

had been segmented by MeCab and merged with the help of a dictionary.

For reading support, we incorporated a system constructed for Japanese language learning that provides English translations for the tokens included in the text as well as phonetic readings (*furigana*) for the *kanji*⁴ characters (Ahltorp, 2012). Topics2Themes was extended to use the *ruby*-tag provided in HTML to display the phonetic reading and one English translation in a small font above each token. In addition, when the user hovers the mouse over a token, all available English translations are shown in the form of a tooltip. To further help the reader, we matched the texts to Japanese sentiment and emotion word lists (Nakamura, 1993; Takamura et al., 2005; Rzepka and Araki, 2012; Rzepka and Araki, 2017), to be able to indicate the existence of such words in the text.

3 Application of the adapted tool on a Japanese corpus

We applied the extended version of Topics2Themes on a corpus consisting of 1,000 microblogs⁵ collected with the criterium that they should contain the same content written in Japanese and in English (Ling et al., 2014). The tool was applied on the Japanese part of the microblogs.

We configured Topics2Themes to try to find 15 topics among the 1,000 texts and to run the topic modelling 100 times, only keeping topics that occurred in all re-runs. This resulted in 12 stable topics being identified by the tool. The most prominent among those can be seen in the *Topics* panel in Figure 1, where each topic is represented by its three most closely associated terms. The small corpus size used, and the small size of each text in the corpus, might make it difficult for the topic modelling algorithm to find reoccurring topics. We therefore configured the tool to allow a large maximum distance for the word2vec-based concept clustering, i.e., two words with a Minkowski distance of up to 0.7 could be included in the same cluster. This makes it possible for the topic model algorithm to find topics based on semantically related words, e.g., on the cluster of food-related words shown in the top element of the *Terms* panel in Figure 1.

Figure 1 also indicates how the results can be explored by the Topics2Themes tool. In the situation shown in the figure, the user has double-clicked on, and thereby selected, the fifth topic in the *Topics* panel. This has had the effect that the terms most closely associated with the selected topic have been sorted as the top-ranked elements in the *Terms* panel, and that the texts most closely associated with the topic have been sorted as the top-ranked elements in the *Texts* panel. The elements associated with the selected topic have also been given a bold, rounded border. The figure also shows how the user hovers the mouse over one of the texts, which has the effect that the terms included in this text, as well as the topic(s) to which the text is associated, are highlighted with a blue colour.

The language support, in the form of phonetic reading and English translation, is shown in a small font above the Japanese texts, as well as in the form of a tool tip for the word over which the mouse hovers. The *Texts* panel also displays the output of the sentiment and emotion word list matching, in the form of labels attached to the texts.

By inspecting the top-ranked terms and texts for each topic, it can be concluded that reoccurring themes in this corpus include greetings, language studies, events in Japan, food, and natural disasters. To further explore recurring themes, the texts selected by the algorithm as most closely associated with the 12 extracted topics should be manually analysed. Such an analysis, using the *Themes* panel for documentation of themes identified, will be included in future work. We intend to let a learner of Japanese perform the analysis, in order to also obtain an indication of the usefulness of the language support provided in our extension of the Topics2Themes tool.

4 Acknowledgements

This research was funded by the Japan Society for the Promotion of Science, and will continue within the Språkbanken and SWE-CLARIN infrastructures, supported by the Swedish Research Council (2017-00626).

⁴The logographic Chinese characters adapted to and used in Japanese.

⁵The corpus used is listed as a CLARIN resource at: <https://www.clarin.eu/resource-families/parallel-corpora>, and is also available at: <http://www.cs.cmu.edu/~lingwang/microtopia/#twittergold>

References

- Magnus Ahltop. 2012. A Personalizable Reading Aid for Second Language Learners of Japanese. Master's thesis, Royal Institute of Technology.
- Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, Oct.
- 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.
- Guoray Cai, Feng Sun, and Yongzhong Sha. 2018. Interactive visualization for topic model curation. In *Joint Proceedings of the ACM IUI 2018 Workshops*. CEUR-WS.
- Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77, New York, NY, USA. ACM.
- Enamul Hoque and Giuseppe Carenini. 2015. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15*, pages 169–180, New York, NY, USA. ACM.
- Chandan K. Jaegul Choo, Haesun Changhyun Lee, Haesun Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1992–2001.
- JMdict. 2013. The JMDict Project. http://www.edrdg.org/jmdict/j_jmdict.html.
- Taku Kudo. 2006. Mecab : Yet another part-of-speech and morphological analyzer. <http://mecab.sourceforge.jp>.
- Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164.
- Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*.
- Wang Ling, Luis Marujo, Chris Dyer, Alan Black, and Isabel Trancoso. 2014. Crowdsourcing high-quality parallel data extraction from twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT '14*, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Akira Nakamura. 1993. *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*. Tokyodo Publishing.
- Rafal Rzepka and Kenji Araki. 2012. Polarization of consequence expressions for an automatic ethical judgment based on moral stages theory. *IPSJ SIG Notes*, 14(2012-NL-207):1–4.
- Rafal Rzepka and Kenji Araki. 2017. What people say? web-based casuistry for artificial morality experiments. In *International Conference on Artificial General Intelligence*, pages 178–187. Springer.
- Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018a. Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, number 247 in Studies in Health Technology and Informatics, pages 366–370. IOS Press.
- Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018b. 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.
- Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Intelligent user interfaces. In *User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System*.
- Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140.

CTS-R: Connecting Canonical Text Services with the Statistical Analytics Environment R

Jochen Tiepmar

Natural Language Processing Group
Department for Computational Humanities
University of Leipzig, Germany
jtiepmar@informatik.uni-leipzig.de

Abstract

This paper describes a software library for the statistical programming language R that builds an interface to the large scale implementation of the Canonical Text Service (CTS) protocol ((Smith, 2009) and (Tiepmar, 2018)). This way the vast amount of textual data that has been and will be collected in the Canonical Text Infrastructure is opened up to all the analytics frameworks and workflows that are available in R. Since the data sets should be usable for any process that is built in R, this drastically increases the reach that these can gain. On the other hand this also increases the amount of textual data that is available in R for textual analysis.

1 Introduction

In (Tiepmar et al., 2016), a newly established workflow to integrate instances of Canonical Text Services (CTS) into CLARIN's research infrastructure was introduced. As this process was repeated in (Grallert et al., 2017), it could be proven that the interface character of the CTS protocol made it possible to implement software solutions that incorporate textual data in a generic way. With the established connection to a persistent archival and data exploration solution in the form of CLARIN, or in especially the Virtual Language Observatory (Goosen and Eckart, 2014), another side of interface connection is still open: The connection to established workflows and frameworks that are in use by current day researchers and not already covered by CLARIN. To respect the generic and universal nature of the work that has been done in the past, this should not be done on individual application level but on the level of programming languages. Two programming languages that are commonly used in the text based digital humanities are Python (van Rossum, 1995) and the more statistics-focused language/framework/environment R (R Core Team, 2014). Since R provides a comfortable framework for web applications and can be assumed to be more common for less technically specialized researchers than Python¹, R can be prioritized.

This paper introduces an open source R package named CTS-R that forms exactly this interface and therefore opens up every piece of text data that is part of the Canonical Text Infrastructure for the vast array of statistical analytical workflows that are part of R. The goal is to allow users of R to implement CTS requests as part of their programs or scripts.

A suitable Python library is currently in development and will be published in the foreseeable future.

2 The R Project for Statistical Computing

R is a programming environment for statistical analysis based on a scripting language. It is especially valuable for humanistic - i.e. social - sciences because it is a free alternative for otherwise relatively expensive software packages like SPSS and its script based nature makes it easy to archive and share analysis workflows. Since its focus is statistical analysis, it can also be used with less detailed programming knowledge and provides comparatively uncomplicated ways to visualize data offline – using core features of R or libraries like ggplot2 – or online – using the package Shiny² – without having to write

¹This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

²Python and its Syntax is more similar to more traditional programming than R.

²<https://shiny.rstudio.com/>

a lot of code. R is especially relevant in the digital humanities because it provides such uncomplicated ways to analyze and visualize data without relying a lot on programming knowledge. Prominent examples of usage would be (Mullen, 2018) or any of the projects listed as part of the world wide R User Groups³.

RStudio⁴ is the software that is recommended by most tutorials in order to work with R. R could be considered as the engine and RStudio as one of its most prominent working environments.

The workflows are distributed using the official Comprehensive R Archive Network CRAN⁵. Additional packages can also be developed and - if compliant - be added to CRAN.

This package is published as a public open source repository but not (yet) as part of CRAN.

3 Package Development and Installation in R

R packages can be developed, tested, documented and installed using the package *devtools*⁶. In short, the package must be compliant with the folder structure illustrated in figure 1 and provide a meta data file *DESCRIPTION* compliant with the one in figure 2. Both illustrations are extracted from the aforementioned cheat sheet.

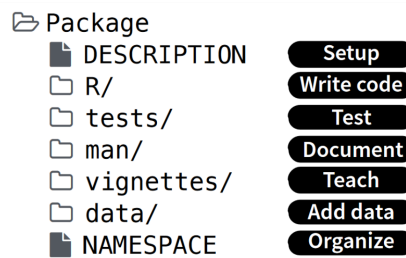


Figure 1: Folder Structure of an R Package

```

Package: mypackage
Title: Title of Package
Version: 0.1.0
Authors@R: person("Hadley", "Wickham", email =
  "hadley@me.com", role = c("aut", "cre"))
Description: What the package does (one paragraph)
Depends: R (>= 3.1.0)
License: GPL-2
LazyData: true
Imports:
  dplyr (>= 0.4.0),
  ggvis (>= 0.2)
Suggests:
  knitr (>= 0.1.0)
  
```

Import packages that your package *must have* to work. R will install them when it installs your package.

Suggest packages that are not very essential to yours. Users can install them manually, or not, as they like.

Figure 2: Content of the DESCRIPTION file

If the package is compliant with the requirements, a user can install a public package that is not part of CRAN using two lines of code as illustrated for the CTS-R package in the following.

³<https://jumpingrivers.github.io/meetingsR/r-user-groups.html>

⁴<https://www.rstudio.com/>

⁵<https://cran.r-project.org/web/packages/policies.html>

⁶A very nice description is published at <https://www.rstudio.com/wp-content/uploads/2015/03/devtools-cheatsheet.pdf>

```
install.packages("devtools")
devtools::install_git("https://git.informatik.uni-leipzig.de/jtiepmar/cts-r/")
```

4 The CTS-R Package

The CTS-R package is built as a wrapper for CTS requests. As such it takes care of the handling of the HTTP requests and XML parsing. That's why it depends on the R packages *httr* and *xml2* that will be installed along with the package.

CTS-R provides 6 main functions that are analogous to their correspondingly named CTS requests: *get_passage_plus*, *get_first_urn*, *get_valid_reff*, *get_prev_next*, *get_label* and *get_passage*. Another function *get_editions* provides document level CTS URNs as returned by the CTS request *GetCapabilities*. Each of the functions returns either the text value of the request or an R vector of elements, so the data can be directly used without any requirement for further parsing or conversion. The following example illustrates both the installation of the package and an example for every function.

```
install.packages("devtools")
devtools::install_git("https://git.informatik.uni-leipzig.de/jtiepmar/cts-r/")

print(ctsRetrieval::get_editions("urn:cts:pbp"))
urn <- "urn:cts:pbp:bible.parallel.eng.kingjames:1"
print(ctsRetrieval::get_prev_next(urn))
print(ctsRetrieval::get_valid_reff(urn,3))
print(ctsRetrieval::get_first_urn(urn))
print(ctsRetrieval::get_passage(urn))
print(ctsRetrieval::get_passage_plus(urn))
print(ctsRetrieval::get_label(urn))
```

As the example illustrates, it is not required to specify a server address that is used for the request. The reason is that this information is automatically added by the package using the Namespace Resolver (Tiepmar, 2018) that maps CTS URNs to specific server addresses. For productive use it should be considered that this is an additional web request for every function use. To avoid redundant requests it is reasonable to use the function *get_namespace_url* to request this server address and provide it for each subsequent request as an optional parameter. An automated approach to this problem using caching would have been possible but was rejected to keep every function self-contained.

5 License

The CTS-R package is published under MIT license. This means that anybody has the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software and this permission notice shall be included in all copies or substantial portions of the Software (Mitchel, 2016).

6 Further work

The CTS implementation provides specialized functions that are not included in this package but could be to add more features and increase performance. Caching techniques could be added to avoid redundant requests. But in general, the CTS-R package should be considered as finalized and connections from CTS to other environments - like Python - should be prioritized to increase the potential reach of the data sets.

An inclusion of this package into CRAN is worth considering. Since submissions to CRAN have high requirements and the usage of this software is possible without it being available as a CRAN package, it is reasonable to wait for initial feedback before doing so.

References

- Twan Goosen and Thomas Eckart. 2014. Virtual Language Observatory 3.0: Whats New?. CLARIN annual conference 2014 in Soesterberg, The Netherlands.
- Kyle E. Mitchell. 2016. The MIT License, Line by Line. <https://writing.kemitchell.com/2016/09/21/MIT-License-Line-by-Line.html>.
- David Neel Smith. 2009. Citation in classical studies. *Digital Humanities Quarterly*, 3.
- Lincoln A. Mullen. 2018. Computational Historical Thinking: With Applications in R. <http://dhr.lincolnmullen.com>.

- Jochen Tiepmar. 2018. Implementation and Evaluation of the Canonical Text Service Protocol as Part of a Research Infrastructure in the Digital Humanities. Dr Thesis, University of Leipzig.
- Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn and Christoph Kuras. 2016. Canonical Text Services in CLARIN - Reaching out to the Digital Classics and beyond. In CLARIN Annual Conference 2016.
- Guido van Rossum. 1995. Python tutorial, Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI), Amsterdam. 1995.
- R Core Team. 2014. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2014.
- Till Grallert, Jochen Tiepmar, Thomas Eckart, Dirk Goldhahn and Christoph Kuras. 2017. Digital Muqtabas CTS Integration in CLARIN. In CLARIN Annual Conference 2017.

Shapeshifting Digital Language Resources – Dissemination Services on ARCHE

Martina Trognitz
ACDH, ÖAW
Vienna, Austria

Matej Ďurčo
ACDH, ÖAW
Vienna, Austria

Abstract

The Austrian Centre for Digital Humanities of the Austrian Academy of Sciences hosts ARCHE – A Resource Centre for the HumanitiEs. ARCHE aims at stable and persistent hosting as well as the dissemination of digital research data and resources for the Austrian humanities community. This paper presents how data in ARCHE can be represented in multiple forms or shapes by using bespoke dissemination services. A focus will be kept on the description of dissemination services for digital language resources, such as XML documents, and showcase a few use cases as well as discuss possible integration of such kind of services into the Virtual Language Observatory of CLARIN.

1 Introduction

ARCHE¹ – A Resource Centre for the HumanitiEs – is the successor of a repository project established in 2014 as the CLARIN Centre Vienna / Language Resources Portal (CCV/LRP). The mission of CCV/LRP was to provide depositing services as well as easy and sustainable access to digital language resources created in Austria. ARCHE replaced CCV/LRP in 2017 and extends its mission by offering a depositing service open to a broader range of humanities fields in Austria. It is hosted by the Austrian Centre for Digital Humanities (ACDH), which is part of the Austrian Academy of Sciences. ARCHE has successfully completed the CLARIN Centre Assessment Procedure and is officially recognised as a CLARIN B Centre.²

ARCHE offers a depositing service for long-term archiving of resources, where each resource can be cited with a persistent identifier. A bespoke metadata schema was created to help in finding, accessing, and reusing data (Trognitz and Ďurčo, 2018). Metadata is required for every resource and all metadata is openly searchable via the web interface. The metadata for language resources is also available in CMDI format, which is harvested by CLARIN in order to allow the resources to be also findable in the VLO.

2 Basic Dissemination Services for Resources in ARCHE

Each resource is presented in a basic user friendly view with selected metadata associated with the resource, as can be seen in figure 1. The metadata section is followed by a list of all dissemination services that are applicable to the resource's type. This growing set of dissemination services³ is intended for the display of specific data types, so that the users can view the resource directly and seamlessly without the need of having to download a resource. A typical example is a TEI file that can be transformed to HTML or PDF for viewing.

Four dissemination services are always shown: "Download", "GUI access", "RDF access", and "Turtle File". The first is intended for the download of the actual binary content of the resource or a collection. The button "GUI access" is basically a self-link to the landing page itself. "RDF access" and "Turtle File" deliver the metadata of a given resource in two alternative serialisation formats.

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

¹<https://arche.acdh.oew.ac.at/>

²<http://hdl.handle.net/11372/DOC-105>

³<https://arche.acdh.oew.ac.at/browser/technical-setup>

In figure 1 two additional dissemination services – ”TEI to DOCX conversion Endpoint” and ”TEI to PDF conversion Endpoint” – are listed, which will be commented on in section 4.

Overview [Copy Resource Link](#) [Switch to Expert-View](#)

● Bestandskontrakt (Pacht) von Kinder des Urban Haas aus Starkenhof (Kreith?) , erstellt 1777_04_05

Type: Resource

Creator(s): Michael Span

Contributor(s): Peter Andorfer

Available Date: 2017-10-16

Extent: 63812

Binary Size: 62.32 KB

License: <https://creativecommons.org/licenses/by-sa/4.0/>

Access Restriction: public

PID: <http://hdl.handle.net/11022/0000-0007-C277-8>

Part of: [privater-buchbesitz/data/editions](#)

Download GUI access RDF access TEI to DOCX conversion Endpoint TEI to PDF conversion Endpoint Turtle File

Figure 1: Basic view of a resource (Span, 2017)⁴ in ARCHE with its dissemination services

3 Binding Dissemination Services to Resources

ARCHE is based on the open-source repository software Fedora Commons (Payette and Lagoze, 1998) version 4 which integrates a triplestore (Blazegraph⁵) for storing the metadata. The dissemination services basically are web applications that are applied to the resources via the so called repo-resolver,⁶ a service to resolve a given resource URI to a particular representation of the resource.

The binding of a dissemination service to a resource can be done on two levels: either on the level of an individual resource or by the application of general matching rules. In both cases metadata is used. The latter case for example asks for the value of a specific property and if that is given, the service will be connected. E. g. a dissemination service can be connected to resources of the MIME type ”text/xml”.

In the first case a property (*hasDissService*) is used to indicate that a resource can be displayed with a specific dissemination service. For example (Hatke, 2017)⁷ is connected to a service to display a given TEI file as HTML via the property *acdh:hasDissService* with the value *TEI2HTML*. Additionally a second metadata property *acdh:hasCustomXSL* with a link to an XSL file is used to allow for custom formatting of the resulting HTML file.

The bindings with the underlying rules are stored as distinct resources in the repository, thus resulting in additional RDF metadata in the underlying triplestore. Therefore they can also be queried for in Blazegraph.⁸

In order to add a new dissemination service it has to be configured first. For each service a title and description is set as well as where the service can be found. The rules to bind a service to resources can also be defined. Any parameters required by the service might additionally be set with a respective property.

Configuration can be done either via a PHP script, which will be processed by dedicated libraries transforming the content into an actual resource in the repository. In listing 1 below an example for a service to transform TEI formatted XML files to PDF with a service provided by OxGarage.⁹

⁵<https://www.blazegraph.com/>

⁶<https://github.com/acdh-oeaw/repo-resolver>

⁷<https://id.acdh.oeaw.ac.at/glasersqueezes2015/rec1110000130/adlib1110000130.xml>

⁸Endpoint: <https://arche.acdh.oeaw.ac.at/blazegraph/#query>

⁹<https://oxgarage2.tei-c.org/>

Another way to configure a dissemination service is by creating an RDF representation and ingesting it as a resource to Fedora. In listing 2 below the resulting RDF from listing 1 is represented in TTL format.¹⁰ Note that rules itself are also stored as resources in Fedora.

Listing 1: PHP code to define a dissemination service

```
// TEI2PDF using OXgarage
$service = new Service($fedora, $idBase . 'TEI2PDF', 'https://arche.
    ↪ acdh.oeaw.ac.at/services/oxgarage/pdf{RES_ID|part(path)}',
    ↪ array('application/pdf', 'pdf'), false);
$service->setMetadata(RF::createMeta(['title' => 'TEI_to_PDF_
    ↪ conversion_Endpoint', RC::get('drupalHasDescription') => '
    ↪ Converts_a_TEI_XML_to_a_PDF']));
$service->addMatch($rdfType, 'http://fedora.info/definitions/v4/
    ↪ repository#Resource', true);
$service->addMatch('http://www.ebu.ch/metadata/ontologies/ebucore/
    ↪ ebucore#hasMimeType', 'application/xml', false);
$service->addMatch('http://www.ebu.ch/metadata/ontologies/ebucore/
    ↪ ebucore#hasMimeType', 'text/xml', false);
$service->updateRms(true, true, $collLoc . '/TEI2PDF');
```

Listing 2: TTL file to define a dissemination service (extract)

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix acdh: <https://vocabs.acdh.oeaw.ac.at/schema#> .
@prefix ldp: <http://www.w3.org/ns/ldp#> .

<https://arche.acdh.oeaw.ac.at/rest/dissServices/TEI2PDF>
  a <https://vocabs.acdh.oeaw.ac.at/schema#DisseminationService> ;
  acdh:hasIdentifier <https://id.acdh.oeaw.ac.at/dissemination/
    ↪ TEI2PDF>, <https://id.acdh.oeaw.ac.at/uuid/93940a2a-7e0f-c3b6
    ↪ -e87f-d424139de33e> ;
  acdh:hasNumberOfItems 4 ;
  acdh:hasTitle "TEI to PDF conversion Endpoint"^^xsd:string ;
  acdh:serviceRevProxy false ;
  acdh:hasDescription "Converts a TEI XML to a PDF"^^xsd:string ;
  acdh:hasReturnType "application/pdf"^^xsd:string,"pdf"^^xsd:string ;
  acdh:serviceLocation "https://arche.acdh.oeaw.ac.at/services/
    ↪ oxgarage/pdf{RES_ID|part(path)}"^^xsd:string ;
  ldp:contains #pointer to binding rules
    <https://arche.acdh.oeaw.ac.at/rest/dissServices/TEI2PDF/_rule1>
    ↪ .
    # this rules hold the matching properties
    # e.g. for MIME type "text/xml":
<https://arche.acdh.oeaw.ac.at/rest/dissServices/TEI2PDF/_rule1>
  acdh:matchesProp http://www.ebu.ch/metadata/ontologies/ebucore/
    ↪ ebucore#hasMimeType;
  acdh:matchesValue text/xml.
```

¹⁰<https://id.acdh.oeaw.ac.at/dissemination/TEI2PDF>

4 Dissemination Services for Language Resources

Currently a total of 18 dissemination services for textual resources, 3d objects, images and georeferenced data is available.¹¹ The services available for annotated textual resources or XML based formats shall be briefly presented here. Where available, example resources where the listed dissemination service can be found and tried out are linked in the respective footnotes.

- OWL2HTML: Allows to view an OWL file as HTML
- CMDI2HTML: Displays CMDI metadata as HTML
- RAW CMDI: View or download raw CMDI metadata¹²
- listPlace2Map: Displays TEI place elements on a map
- customTEI2HTML: Transforms a TEI document with a custom XSLT stylesheet into HTML¹³
- TEI2HTML using OxGarage: Uses OxGarage to display a TEI encoded XML as HTML
- TEI2PDF using OxGarage: Uses OxGarage to provide a TEI encoded XML as PDF¹⁴
- TEI2DOCX using OxGarage: Uses OxGarage to provide a TEI encoded XML as DOCX for download¹⁵

All these services run as stand alone applications, which also allows them to be used as resource type specific viewers by third parties. For example they could be integrated in VLO to preview the discovered resources, as additional option to just pointing to their original location. Alternatively dissemination services could also be provided to VLO via a bespoke metadata property, because the services are simply used by calling their URL and attaching the resource's UUID to it, as e. g. <https://arche.acdh.oeaw.ac.at/services/oxgarage/pdf/uuid/3a880b61-3ae1-e16f-c345-339ce60ab4c5>.

5 Conclusion and Outlook

Dissemination services provide a means to add further functionality to a repository beyond simple viewing of metadata and downloading resources. By using different services a single file can be provided in several formats. In this way for example a single TEI encoded file can be viewed online in HTML or downloaded in DOCX or PDF format.

ARCHE's dissemination services not only allow to provide different formats for a single file via the basic view of a resource in the repository. For what is more, the method of attaching the resource's UUID to the URL of the dissemination service can serve the same resource in a different application without the need of storage duplication.

References

- George Hatke. 2017. Digital Edition of Glaser Abklatsch ID: AT-OeAW-BA-3-27-A-GL1000b.01rr.
- Sandra Payette and Carl Lagoze. 1998. Flexible and extensible digital object and repository architecture (fedora). In Christos Nikolaou and Constantine Stephanidis, editors, *Research and Advanced Technology for Digital Libraries*, pages 41–59, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Michael Span. 2017. Bestandskontrakt (Pacht) von Kinder des Urban Haas aus Starkenhof (Kreith?) , erstellt 1777.04.05.
- Martina Trognitz and Matej Ďurčo. 2018. Ein Schema, sie alle zu binden. Das Innenleben des digitalen Archivs ARCHE. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 71(1):217, July.

¹¹<https://arche.acdh.oeaw.ac.at/browser/discover/&type=DisseminationService>

¹²<http://hdl.handle.net/11022/0000-0001-8B04-E>

¹³<https://id.acdh.oeaw.ac.at/glasersqueezes2015/rec1110002525/adlib1110002525.xml>

¹⁴<http://hdl.handle.net/11022/0000-0007-C277-8>

¹⁵<http://hdl.handle.net/11022/0000-0007-C277-8>

Wablief: An Easy-to-Read Newspaper Corpus for Dutch

Vincent Vandeghinste

Instituut voor de Nederlandse Taal

Leiden, the Netherlands

vincent.vandeghinste@ivdnt.org

Bram Bulté

KU Leuven

Belgium

bult@ccl.kuleuven.be

Liesbeth Augustinus

KU Leuven

Belgium

liesbeth@ccl.kuleuven.be

Abstract

This paper presents the Wablief corpus, a two million words corpus of a Belgian easy-to-read newspaper, written in Dutch. The corpus was automatically annotated with CLARIN tools and is made available in several formats for download and online querying, through the CLARIN infrastructure. Annotations consist of part-of-speech tagging, chunking, dependency parsing, named entity recognition, morphological analysis and universal dependencies. By making this corpus available we want to stimulate research into text readability and automated text simplification.

1 Introduction

Easy-to-read texts are texts that are targeted at people with a limited functional literacy. According to the United Nations Handbook of Household Surveys, "*a person is functionally illiterate who cannot engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing and calculation for his own and the community's development*" (United Nations, 1984). Limited functional literacy is not an infrequent phenomenon. One out of ten people in Flanders, Belgium are low literate, and have trouble reading text on paper and on web sites. Many texts are too difficult, for example because they contain too many difficult words and/or complex sentences.

The goal of the Wablief organisation is to address this issue on both sides: people who like to read easy texts can read the Wablief newspaper, and people and organisations that want to publish easily readable texts can take training sessions at Wablief, or can ask Wablief to rewrite their texts. The Wablief newspaper,¹ established in 1989, is a weekly newspaper in so-called *clear* language ("duidelijke taal" in Dutch), with more than 10,000 readers. An online archive of volumes published since 2009 is available on the organisation's website, and it is this online archive which is now made available as an automatically linguistically annotated corpus of about two million word tokens.

Our aim is to make this corpus available to the natural language processing community as a target corpus of clear, easy-to-read Dutch for automatic simplification tools, to the applied linguistics community as an example of texts that were written with the explicit intention of being accessible, even by readers with low literacy skills, and to the linguistics community in general as a set of texts showing the linguistic characteristics of what is understood as *clear* writing in Flanders. The corpus has already been used by Bulté et al. (2018) in a lexical simplification task for the identification of difficult words and the selection of potentially easier replacements.

2 Related work

There are a number of approaches towards writing text while addressing the issue of limited literacy. Some texts are written with the general aim of being *easy-to-read*, others have children as their target group, and yet other texts aim at people with cognitive disabilities, so there is a large spectrum of possible target users.

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: [http://creativecommons.org/licenses/by/4.0/finalpaper:en-ukversion\(tolicense,alicence\)](http://creativecommons.org/licenses/by/4.0/finalpaper:en-ukversion(tolicense,alicence))

¹<http://www.wablief.be/>

Probably the most well-known initiative in this respect is Wikipedia Simple English,² which is written in so-called *simple English*. Amongst other things, its authors are instructed to use only the 1000 most frequent words of English. Other Wiki-initiatives are Wikikids³ for Dutch-speaking children and Wikidia⁴ for speakers of French, Italian, Spanish, English, Basque, Catalan, German, Russian, Greek and Sicilian.

In an academic context, a number of available easy-to-read corpora have been described and used, such as the Swedish LäsBarT corpus (Mühlenbock, 2009), a corpus for Brazilian Portuguese (Aluísio et al., 2008), the French CLEAR medical corpus (Grabar and Cardon, 2018), and a very small (227 sentences) corpus for Basque (Gonzalez-Dios et al., 2018). There have been some efforts to compile monolingual comparable corpora, aligning *normal* text with its *easy-to-read* variant. Alignment can be at the text level, the paragraph level or the sentence level. A list of English comparable corpora with an easy-to-read side can be found in Yaneva (2015), and also for French (Cardon and Grabar, 2018) and Brazilian Portuguese (de Medeiros Caseli et al., 2009) there have been efforts to create such a comparable corpus.

We are not aware of any such efforts for Dutch. Concerning more simple forms of Dutch, the Dutch data in the CHILDES project might be worth mentioning (MacWhinney, 2000), as well as the JASMIN speech corpus, consisting of recordings of Dutch speech by young people, non-native speakers, and elderly people (Cucchiari et al., 2008). These two projects record *active* speech, whereas the Wabliet corpus contains texts focusing on the *passive* language knowledge of the target users.

3 Corpus processing and availability

3.1 Creation of the metadata file

We received the data from our data providers as a zip file containing a number of text files, with no further information. The names of the text files were structured according to the regular expression pattern in (1), with the first set of digits indicating the newspaper volume (between the first set of parentheses), the article category (between the second set of parentheses), and the article number inside this category (between the third set of parentheses).

$$/wa(\d{3,4})(bi|ka|\dots)(\d).txt/ \quad (1)$$

The dates from the text files we received mostly agree with the publication dates, so we took this information as the publication dates of the articles in the metadata file, which we provide as a tab-separated value file (tsv). Table 1 presents the different categories that are distinguished in the newspaper and in the corpus.⁵

3.2 Automated annotations

We used the LaMachine unified software distribution for Natural Language Processing⁶ to perform processing with Frog (Van den Bosch et al., 2007) and with Alpino (van Noord, 2006). LaMachine was available through the CLARIN Switchboard. Unfortunately, this is no longer the case, due to the separate registration procedure for the servers of the Radboud University in Nijmegen.⁷

Frog is an NLP suite based on memory-based learning and trained on large quantities of manually annotated data. It automatically annotates the word tokens in Dutch text files. Frog's output is available in two formats, which contain the same information: the FoLiA format (van Gompel and Reynaert, 2013), and the tab-delimited column-formatted output, one line per token (also known as CoNLL format). The ten columns contain (1) the token number within the sentence; (2) the token itself; (3) the predicted lemma; (4) the predicted morphological segmentation; (5) the predicted part of speech (PoS) tag; (6) the confidence with which the PoS tag was predicted; (7) the predicted named entity type, distinguishing

²https://simple.wikipedia.org/wiki/Main_Page

³<http://www.wikikids.nl>

⁴<http://www.wikidia.org>

⁵The presented numbers are those of the treebank version and might slightly differ from the non-treebank versions, as parsing might have failed in a number of cases.

⁶<https://proycon.github.io/LaMachine/>

⁷<https://webservices-1st.science.ru.nl/register>

Dutch name	English name	number of sentences	number of words
Binnenland	Domestic	58,560	489,296
Blog	Blog	5,464	40,990
Buitenland	Foreign	40,953	337,536
Cijfer van de week	Number of the week	1,595	11,973
In de kijker	In the spotlight	49,438	398,366
Jaaroverzicht	Annual overview	301	2,484
Mening	Opinion	6,117	45,860
Samenleving	Society	23,660	189,847
Sport	Sports	24,747	196,697
Tip	Hint	12,869	100,513
Verhaal	Story	2,529	19,581
Voorpagina	Front page	5,121	42,527
Weetjes	Facts	19,927	156,222
Zomer	Summer	5,448	42,599
Total		256,729	2,074,491

Table 1: Categories of the Wablieft corpus

between *person*, *organization*, *location*, *product*, *event*, and *miscellaneous*, using a IOB encoding;⁸ (8) the predicted phrase chunk in BIO encoding; (9) the predicted token number of the head word in the dependency graph; and (10) the predicted type of dependency relation with head word.

Alpino is a hybrid dependency parser for Dutch, which uses rule-based constraints combined with corpus-based statistics. It provides its own XML tree format, which is isomorphic to the syntax tree, unlike XML tree representations like FoLiA and TigerXML (Lezius et al., 2002). This makes it suitable for XPath and XQuery searches and scripts, allowing easy inclusion in CLARIN treebank query tools like GrETEL (Augustinus et al., 2017) and PaQu (Odijk et al., 2017). We also provide a Universal Dependencies annotation⁹ in CoNLL-UD format, in which the Alpino parses are automatically converted into CoNLL-UD using the script from Bouma and van Noord (2017).¹⁰

3.3 Availability

The corpus can be downloaded for non-commercial purposes at the Dutch Language Institute, the CLARIN-B centre for Flanders.¹¹ It comes with several (automatic) annotations and is delivered in a variety of formats: one directory per newspaper article, with one file per sentence in alpino XML; one XML file per newspaper article; frequency lists; the *frogged* versions of the files with automatic sentence detection, both in FoLiA and CoNLL format; the newspaper articles, one sentence per line (automatic detection); the original texts with paragraph and header markup, fixed for UTF-8; and one file with the Universal Dependencies annotation in CoNLL-UD format. We also included Wablieft in the GrETEL treebank query tool (Augustinus et al., 2017),¹² so it can easily be queried.

4 Conclusions and future work

We presented the Wablieft corpus and how it was automatically annotated using the CLARIN infrastructure. The corpus is also made available through the CLARIN infrastructure.

In future work, we want to create a comparable corpus through aligning articles from regular newspapers with the articles in the Wablieft corpus, not unlike Cardon and Grabar (2018), on which we can train a text simplifier for Dutch. We also intend to use this corpus for language modelling of easy-to-read language.

⁸[https://en.wikipedia.org/wiki/Inside-outside-beginning_\(tagging\)](https://en.wikipedia.org/wiki/Inside-outside-beginning_(tagging))

⁹<https://universaldependencies.org/>

¹⁰<https://github.com/gossebouma/lassy2ud>

¹¹<https://ivdnt.org/downloads/taalmaterialen/tstc-wablieft-corpus-1-1>

¹²<http://gretel.ccl.kuleuven.be/>

References

- Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008. A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, pages 15–22, New York, NY, USA. ACM.
- Liesbeth Augustinus, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde. 2017. GrETEL: A Tool for Example-based Treebank Mining. In *CLARIN in the Low Countries*, chapter 22, pages 269–280. London: Ubiquity Press.
- Gosse Bouma and Gertjan van Noord. 2017. Increasing return on annotation investment: the automatic construction of a Universal Dependency treebank for Dutch. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 19–26, Gothenburg, Sweden.
- Bram Bulté, Leen Sevens, and Vincent Vandeghinste. 2018. Automating lexical simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, 8:24–48, Dec.
- Rémi Cardon and Natalia Grabar. 2018. Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 83–93. Association for Computational Linguistics.
- Catia Cucchiarini, Joris Driesen, Hugo Van hamme, and Eric Sanders. 2008. Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In *LREC 2008*.
- Helena de Medeiros Caseli, Tiago de Freitas Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline, Gasperin, and Sandra M. Aluísio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *Proceedings of CICLing*.
- Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2018. The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation*, 52(1):217–247, Mar.
- Natalia Grabar and Rémi Cardon. 2018. Clear – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9. Association for Computational Linguistics.
- Wolfgang Lezius, H. Biesinger, and Ciprian Gerstenberger, 2002. *Tiger-XML quick reference guide*.
- Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.
- Katarina Mühlenbock. 2009. Readable, legible or plain words - presentation of an easy-to-read Swedish corpus. In *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics*, volume Studia Linguistica Upsaliensia 8, pages 325–327.
- Jan Odijk, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang. 2017. The parse and query (PaQu) application. In *CLARIN in the Low Countries*, chapter 23, pages 281–297. London: Ubiquity Press.
- United Nations. 1984. *Handbook of Household Surveys, Revised Edition*, volume No. 31 of *Studies in Methods, Series F*. United Nations, New York.
- Antal Van den Bosch, Gert Jan Busser, Walter Daelemans, and Sander Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium.
- Maarten van Gompel and Martin Reynaert. 2013. Folia: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, 3:63–81, Dec.
- Gertjan van Noord. 2006. At last parsing is now operational. In *TALN 2006*, pages 20–42.
- Victoria Yaneva. 2015. Easy-read documents as a gold standard or evaluation of text simplification output. In *Proceedings of the Student Research Workshop*, pages 30–36. INCOMA Ltd. Shoumen, Bulgaria.