CLARIN-D Report R8.5

Hannah Kermes, Jörg Knappen, José Manuel Martínez Martínez, Elke Têich, Mihaela Vela

May 2016
1 Introduction

The major event of work package Training and Education in the fifth project year was the joint summer school with the European summer university Digital Humanities Culture & Technology (ESU DH C&T) which took place from July 28th to August 7th in Leipzig. CLARIN-D co-organized the summer university for the second time, organized three workshops and an afternoon project presentation session explaining and showing

- what you can do with CLARIN-D tools and resources,
- how to process your own resources and integrate them in CLARIN-D and
- how to contact and/or get help from CLARIN-D.

Both the workshops and the project presentation session were well received. A more detailed report on the summer university with a list of the workshops and presentations can be found in chapter 2 and on the website of the summer university (1).

The next joint venture with the ESU DH C&T is already in planning: The third joint summer school will take place from July 19th to July 29th, again in Leipzig. CLARIN-D will again organize five workshops. More information on the workshops and on the status of the summer university can be found in chapter 3 and on the website of the summer university (2).

Apart from the summer university, CLARIN-D organized a number of individual workshops, tutorials, courses, demonstrations and presentations at various events and locations including a workshop series called Working with Corpora, which followed a concept of learning from one another. An overview of these training measures can be found in chapter 4.

For TeLeMaCo, the collaborative portal for teaching and learning material, we saw a steady growth in teaching material. The newly implemented access monitoring shows approximately 300 accesses per month. For more information on TeLeMaCo see chapter 5.

1 http://www.culingtec.uni-leipzig.de/ESU_C_T/
2 http://www.culingtec.uni-leipzig.de/ESU_C_T/
2 Second Joint CLARIN-D and European Summer University
Digital Humanities Culture & Technology

The second joint CLARIN-D and European Summer University in Digital Humanities Culture & Technology took place from July 28th to August 7th at the Geisteswissenschaftliches Zentrum of the Universität Leipzig. You can find more information on the general concept and the mission of the summer university on the website and in project report R8.3.

The structure of the summer university was similar to that of its predecessor. Centered around a selection of small workshops (aiming at a maximum of 10 participants), there were plenaries including lectures, presentations, poster sessions, and a round table. Please consult the website for a complete schedule of the summer university.

2.1 CLARIN-D workshops

CLARIN-D organized three workshops:

- Axel Herold (Berlin-Brandenburgische Akademie der Wissenschaften, Germany) / Erhard Hinrichs (University of Tübingen, Germany) / Thorsten Trippel (University of Tübingen, Germany): Methods and Tools for the Corpus Annotation of Historical and Contemporary Written Texts
- Peter Fankhauser (Institut für Deutsche Sprache, Mannheim, Germany) / Hannah Kermeres (Saarland University, Germany) / Elke Teich (Saarland University, Germany): Comparing Corpora
- Matthias Lang / Dieta Svoboda (University of Tübingen, Germany): Spatial Analysis in the Humanities

All workshops were so popular that they were overbooked very fast and more than 10 participants were accepted in the end.

As a conclusion, we can say that this second edition of the joint summer school showed an increase in popularity of CLARIN-D and the resources, tools and services it has to offer among the participants of the summer university and also among the larger DH community. Many students and young researchers had been sent by their supervisors to CLARIN-D courses to learn how to use the CLARIN-D infrastructures and the possibilities it offers.

2.2 CLARIN-D Projects session

A monographic session devoted to the CLARIN-D project took place on 29 July 2015 during the joint ESU/CLARIN-D summer school. The aim of the session was to foster curiosity.
among participants and illustrate the possibilities of language technology applied to different problems. Support resources like the CLARIN-D Helpdesk, the User Guide, and the Teaching and Learning Materials Collection were also presented.

The session was very successful. The conference room at the Bibliotheca Albertina was completely full. The session had gained a lot of interest. The reception was very positive, and the session worked both as a stimulating kick-off and as an ice-breaker leading to fruitful discussions in the following days between participants and CLARIN-D members present at the event.

The session was moderated by José Manuel Martínez Martínez (Saarland University). After a brief introduction to CLARIN-D, its the goals and the main components of the infrastructure by Thorsten Trippel (University of Tübingen), several projects and resources were presented on the following topics:

- What can you do with CLARIN-D tools and resources?
  - *Textual Emigration Analysis* (Wiltrud Kessler and André Blessing, Institute for Natural Language Processing, University of Stuttgart)

- How can you process your own resources and integrate them into CLARIN-D?
  - *Applying Optical Character Recognition* (Axel Herold, Berlin-Brandenburg Academy of Sciences and Humanities, Berlin)
  - *EXMARaLDA + WebLicht as a Service* (Hanna Hedeland, Hamburg Center for Language Corpora, University of Hamburg)
  - *COALA Webservice* (Bernhard Jackl, Bavarian Archive for Speech Signals, Ludwig-Maximilian University of Munich)

- How can you contact and/or get help from CLARIN-D?
  - *CLARIN-D Helpdesk* (Timm Lehmberg, Hamburg Center for Language Corpora, University of Hamburg)
  - *TeLeMaCo* (José Manuel Martínez Martínez, English Linguistics and Translation Science, Saarland University)
3 Third Joint CLARIN-D European Summer University in Digital Humanities Culture & Technology

The third joint CLARIN-D European Summer University in Digital Humanities Culture & Technology will take place from July 19th to July 29th at the Geisteswissenschaftliches Zentrum in Leipzig. We decided to continue the successful partnership between CLARIN-D and the ESU DH C&T by co-organizing a third event. The structure of the summer university will in principal remain the same: small hands-on workshops with up to 10 participants, plenaries including lectures, presentations, a poster session, and a round table. The number of total participants is again aimed at around sixty. CLARIN-D will organize five workshops:

- Christoph Kuras, Gregor Wiedemann (both Leipzig): Lexicometric text analysis using CLARIN-D Webservices and R. (1 week)
- Alex Herold, Susanne Haaf (both BBAW): Compilation, Annotation und Analysis of Written Text Corpora. Introduction to Methods and Tools. (1 week)
- Erhard Hinrichs (TüB), Bryan Jurish, Alexander Geyken, Lothar Lemnitzer (all BBAW): Searching Linguistic Patterns in Large Text Corpora for Digital Humanities Research. (1 week)
- Pawel Kamocki (IDS), Thorsten Trippel (Tüb): Data management for the humanities: from data warehousing to legal and ethical implications. (2 weeks)
- Peter Fankhauser (IDS), Hannah Kermes, Elke Teich (both UdS): Comparing Corpora. (1 week)

Please visit the website of the summer university for a full list of workshops.

Aside of financing parts of the costs of the summer school, CLARIN-D will again give out scholarships covering tuition fees.

[http://www.culingtec.uni-leipzig.de/ESU_C_T/node/621](http://www.culingtec.uni-leipzig.de/ESU_C_T/node/621)
4 Training modules and tutorials

4.1 Workshops, tutorials and courses

4.1.1 Course: Transkription I & II

Timm Lehmbberg (HZSK)
Seminar “Sprachpolitik”, Universität Hamburg
2 & 16 June 2015, Hamburg
The course featured an introduction to transcription and annotation using the EXMARaLDA software, which was integrated in two parts into a seminar on language policies and politics.

4.1.2 Tutorial: Search for Needles in DH Haystacks Continued: Zooming in Corpus Query Tools

Andreas Witt, Piotr Bański, Nils Diewald, Michael Hanl, Marc Kupietz (IDS)
Digital Humanities 2015
29 June 2015, Sydney, Australia
The tutorial presented the state-of-the-art methods in querying textual data with use cases commonly found in digital humanities, or envisioned for the near future. It demonstrated how to carry over some well-known methods and techniques from linguistic research to a broader area of digital humanities. The tutorial focused primarily on the search in metadata, non-annotated data, and structured annotated data (especially TEI-encoded), and showed how to ‘zoom in’ on text collections to discover or confirm the existence of interesting data patterns.

4.1.3 Workshop: Angebote von CLARIN-D für die Romanistik

Dr. Dörte de Kok, Wiltrud Kessler, José Manuel Martinez Martinez, Pawel Kamocki (CLARIN-D)
XXXIV Romanistentag
27 July 2015, Mannheim
http://www.romanistentag.de/index.php?id=860
In this workshop, CLARIN-D introduced the tools and services it offers. The main focus was on those tools and services directly relevant for Romance research. The presentations included information on data management, Romance resources, tools for finding and analyzing Romance data and legal issues.

4.1.4 Workshop: Methods and Tools for the Corpus Annotation of Historical and Contemporary Written Texts

Axel Herold (BBAW), Erhard Hinrichs (UTüb), Thorsten Trippel (UTüb)
This course dealt with methods and tools for the annotation of corpora. The main focus were written texts, of both historical and contemporary origin. Participants learned to use various existing data categories and tools for manual and automatic annotation as well as ways of using annotations for research questions.

4.1.5 Workshop: Special Session “Basic audio and video recording techniques, (meta-)data formats and ethical and legal aspects”

Hanna Hedeland (HZSK), Bernhard Jackl (BAS), Timm Lehmberg (HZSK)
Special Session at the ESU Summerschool
30 July 2015, Universität Leipzig
The special session on speech and spoken language corpora covered practical aspects regarding the recording and creation of corpora, standards and recommended formats, and fundamental questions regarding ethical and legal aspects.

4.1.6 Course: Comparing Corpora

Peter Fankhauser (IDS), Hannah Kermes, Elke Teich (both Saarland University)
ESU European Summer University on Digital Humanities Leipzig 2015
3–7 August 2015, Leipzig
The course covered different aspects of corpus analysis regarding the comparison of corpora or sub-corpora in the area of register analysis, diachronic analysis, and contrastive analysis (mostly German English). It discussed different methods ranging from traditional corpus linguistic analysis to data mining and machine learning techniques, from data exploration to data analysis and interpretation.

4.1.7 Workshop: New digital language resources and infrastructure for English (Language) Studies

Dr. Thorsten Trippel, Prof. Dr. Christian Mair
Anglistentag 2015
23 September 2015, Paderborn
Workshop on CLARIN-D, its goals and the main components of the infrastructure.

4.1.8 Course: Corpus Query with Query Language: XQuery and specialized corpus languages

Andreas Witt (IDS)
Digital Humanities Autumn School 2015 “Methodological Interchapters”, University of Luxembourg, Digital History Lab
2 October 2015, Esch-sur-Alzette, Luxembourg
The course presented different methods in querying textual data focusing on regular expressions, XQuery and some query languages used in linguistics such as TEI.
4.1.9 **Workshop: Linguistische Annotationstools als Web-Applikation: das Beispiel WebLicht**

Marie Hinrichs (UTüb), Dr. Claus Zinn

ACDH Tool Gallery 1.2: Basic Text Enrichment—TreeTagger for DH-Applications
13 October 2015, Vienna, Austria

[http://www.oeaw.ac.at/acdh/en/acdh-tool-gallery-1.2](http://www.oeaw.ac.at/acdh/en/acdh-tool-gallery-1.2)

WebLicht is a web application (research environment) for the analysis and annotation of text. Through CLARIN, WebLicht is available for all researchers and integrates a large number of linguistic tools (tokenizer, tagger, parser) for different languages. In this session, the basic handling of this application will be introduced and practiced with exemplar texts.

4.1.10 **Course: EXMARAlda**

Timm Lehmberg (HZSK)

Seminar “Methoden empirischen Arbeitens”, Universität Hamburg
18 November 2015, Hamburg

The course featured a short introduction to transcription and annotation using the EXMARAlda software and was given to students in a seminar on empirical methods.

4.1.11 **Course: Einführung in das Transkribieren mit EXMARAlda**

Timm Lehmberg (HZSK)

EXMARAlda Training, Universität Hamburg
27 November 2015, Hamburg


The one-day course featured a thorough introduction to transcription and annotation using the EXMARAlda software and also treated corpus management and analysis. The course was open to students and employees of the Hamburg University and other universities.

4.1.12 **Tutorial: ICARUS**

Markus Gärtner (IMS, University of Stuttgart)

Seminar “Computer Skills for Computational Linguists”
Winter term 2015/2016, Stuttgart

Practical Introduction to ICARUS split across 3 sessions

ICARUS is an interactive tool for exploration and search of corpora. Being a representative for modern utility tools in NLP and DH, it was used in this course to demonstrate how such tools can help in answering research questions and how to get a better understanding of a given corpus resource. Students learned the capabilities (and limits) of high level corpus search tools and the usage of the ICARUS tool in particular.

4.1.13 **Course: CLARIN webservice (students’ project)**

André Blessing, Markus Gärtner, Jonas Kuhn, Nils Reiter (IMS, University of Stuttgart)

Students’ project
Winter term 2015/2016, Stuttgart
This hands-on course first introduces into existing software solutions for real-world (computational) linguistics questions and difficulties. One project of a students’ group (“FictionNet”) was to implement an interface to WebLicht to create a web-based platform for analyzing drama texts.


Christian Thomas, Susanne Haaf, Matthias Boenig (all BBAW)
Workshop series of DHnet Jena
15–16 January 2016, Jena
http://dhnet.uni-jena.de/index.php?id=124
http://www.deutschestextarchiv.de/news/52
The workshop was focused on methods for the creation and analysis of language corpora. Participants were introduced in the work with the Deutsches Textarchiv corpora as well as with the offerings of CLARIN-D.

4.1.15 Course: EXMARaLDA

Timm Lehmberg (HZSK)
Seminar “Einführung in die Linguistik des Deutschen”, Universität Hamburg
16 Februar 2016, Hamburg
The course featured a short introduction to transcription and annotation using the EXMARaLDA software for students in an introductory course on German Linguistics.

4.2 Demos

4.2.1 Demo: EXMARaLDA and WebLicht as a Service

Hanna Hedeland (HZSK)
CLARIN-D Project session at the ESU Summerschool
29 July 2015, Leipzig
A demonstration of the new integration of WebLicht as a Service into the EXMARaLDA transcription and annotation tool within the CLARIN-D Project session of the summer university.

4.2.2 Demonstration: LINDAT License Selector

Pawel Kamocki (IDS), Pavel Stranak (Charles University)
CLARIN Annual Conference 2015, Bazaar session
16 October, 2015, Wroclaw, Poland
Demonstration of the LINDAT License Selector (developed by Pawel Kamocki, Pavel Stranak and Michal Sedlak) during the CLARIN Bazaar session. The tool is intended to guide users through the choice of a licence for their software and data.
4.2.3 Demo: Presentation of WebLicht and VLO

Jens Stegmann (IMS, University of Stuttgart)
Course in Text Technology
Winter term 2015/2016, Stuttgart
One aspect of this course was to introduce students to foundational XML technologies (documents, DTDs, XML Schema, XSLT/Xpath) used for processing and interchanging documents. The course also covered linguistic annotation, different types of pertinent resources, related standardization efforts and demonstrations of CLARIN-specific applications, e.g., WebLicht and the VLO.

4.2.4 Demo: CLARIN Booth at the DHd 2015

Dr. Thorsten Trippel (UTüb)
Digital Humanities deutschsprachiger Raum
7–12 March 2016, Leipzig
http://www.dhd2016.de
Digital Humanities in the German speaking area is a conference of the humanities scholars interested in using digital solutions. Participants are potential CLARIN users already in testing phases, as they are experienced with services in testing state and can provide valuable feedback while applying the infrastructure. The conference booth was used for special sessions, hands on training, answering questions and establishing contact.

4.2.5 Demonstration: Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP, der neuen Korpusanalyseplattform des IDS

Marc Kupietz, Nils Diewald, Michael Hanl, Eliza Margaretha (IDS)
IDS Jahrestagung 2016
9 March 2016, Mannheim, Germany
A demonstration of the KorAP (corpus analysis platform) system guiding users how to use the system and presenting possibilities of performing complex linguistic query to search in DeReKo.

4.2.6 Demonstration: The Public License Selector: Making Open Licensing Easier

Pawel Kamocki (IDS), Pavel Stranak (Charles University)
LREC Main Conference 2016
25–27 May 2016, Portoroz, Slovenia
A demonstration of the LINDAT License Selector (developed by Pawel Kamocki, Pavel Stranak and Michal Sedlak) intended to guide users through the choice of a licence for their software and data.

4.3 Presentations

4.3.1 Presentation: Mehrsprachige Korpora und digitale Forschungsinfrastrukturen - das Hamburger Zentrum für Sprachkorpora

Hanna Hedeland, Timm Lehmborg (HZSK)
Forschungskolloquium “eHumanities”, Universität Hamburg
18 June 2015, Hamburg
A presentation of the CLARIN-D resources and infrastructure at the HZSK within the research colloquium “eHumanities” of the University of Hamburg.

4.3.2 Presentation: Language Resource Management—Corpus Query Lingua Franca (CQLF) Part 1: Metamodel

Andreas Witt (IDS)
Annual meeting the ISO Working Group TCF37/SC4 Language Resources, 23 June 2015, Matsue, Japan
The presentation described the CQLF standard development.

4.3.3 Presentation: CLARIN-D – Ziele, aktueller Stand, Use cases der digitalen Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Prof. Dr. Erhard Hinrichs, Prof. Dr. Gerhard Heyer
Workshop “Fächerübergreifende Perspektiven durch digitale Forschungsinfrastrukturen” 30 June–1 July 2015, Leipzig
http://clarin2015.informatik.uni-leipzig.de
In this presentation CLARIN-D describes the goals and the current state of the infrastructure. Use cases play an important role in providing examples for the use of the infrastructure.

4.3.4 Presentation: Depositing Services: Wie bekomme ich meine Ressourcen in die CLARIN-D Infrastruktur

Dr. Thorsten Trippel
Workshop “Fächerübergreifende Perspektiven durch digitale Forschungsinfrastrukturen” 30 June–1 July 2015, Leipzig
http://clarin2015.informatik.uni-leipzig.de
In this presentation the new CLARIN-D data management wizard was introduced. Furthermore the depositing service for external data was discussed in detail.

4.3.5 Presentation: Mehrsprachige Korpora und digitale Forschungsinfrastrukturen - das Hamburger Zentrum für Sprachkorpora

Hanna Hedeland, Timm Lehmbeg (HZSK)
Ringvorlesung “Korpulsinguistik”, Universität Bremen 8 July 2015, Bremen
A presentation of the CLARIN-D resources and infrastructure at the HZSK within the lecture series “Korpulsinguistik” of the University of Bremen.

4.3.6 Poster presentation: An Open Source Corpus-Query Platform for the Analysis of Very Large Multiply Annotated Corpora

Piotr Bański, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Eliza Margaretha, Andreas Witt (IDS)
3rd Workshop on the Challenges in the Management of Large Corpora, Corpus Linguistics Conference, Lancaster University
20 July 2015, Lancaster, UK
The poster presents an overview of the KorAP system including its data model and various components (e.g. Kalamar user front-end, Koral query serializer, Krill Lucene-based search engine, Kustvakt user and policy management). KorAP’s data model makes a separation between corpus text and its annotations allowing unlimited addition of annotation layers.

4.3.7 Presentation: Legal Issues in Digital Humanities

Erik Ketzan (IDS)
European Summer School in Digital Humanities, Universität Leipzig
29 July 2015, Leipzig

4.3.8 Presentation: CLARIN-D Helpdesk

Timm Lehmburg (HZSK)
CLARIN-D Project session at the ESU Summerschool, Universität Leipzig
29 July 2015, Leipzig
A presentation of the CLARIN-D Helpdesk and the network of expertise behind it available to users of the CLARIN-D infrastructure.

4.3.9 Presentation: 'All Your Data are Belong to Us'. European Perspective on Privacy Issues in ‘Free’ Online Machine Translation Service

Pawel Kamocki (IDS), Jim O’Regan (Trinity College) and Marc Stauch (University of Hanover)
10th International IFIP Summer School
19 August 2015, Edinburgh, UK
Presentation at the 10th IFIP (International Federation for Information Processing) Summer School in Edinburgh. It discusses how free on-line MT services could fit within the EU data protection framework. The paper, co-authored by Pawel Kamocki, Jim O’Regan (MT developer from Trinity College) and Dr. Marc Stauch (an expert on privacy law from the University of Hanover) has been accepted for publication in conference proceedings (to appear later in 2016).

4.3.10 Presentation: Vom digitalen Hausmeister zum international vernetzten Forschungsdatenzentrum. Die Entwicklung der digitalen Infrastruktur des Hamburg Zentrums für Sprachkorpora

Daniel Jettka, Hanna Hedeland (HZSK)
Forschungsdaten in den Geisteswissenschaften (FORGE 2015), Universität Hamburg
17 September 2015, Hamburg
[https://lecture2go.uni-hamburg.de/veranstaltungen/-/v/18297](https://lecture2go.uni-hamburg.de/veranstaltungen/-/v/18297)
A presentation of the development of CLARIN-D compatibility for the resources and infrastructure at the HZSK.
4.3.11 Presentation: INEL – ein Infrastrukturvorhaben zur Dokumentation indigener nordeurasischer Sprachen

Beáta Wagner-Nagy (IFUU/UHH), Timm Lehmberg (HZSK)
Forschungsdaten in den Geisteswissenschaften (FORGE 2015), Universität Hamburg
18 September 2015, Hamburg
https://lecture2go.uni-hamburg.de/veranstaltungen/-/v/18306

A presentation of the project INEL (Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages) and its integration into the HZSK and CLARIN-D infrastructure.

4.3.12 Presentation: Wissenstransfer und Wissensressourcen: Support und Helpdesk in den Digital Humanities

Timm Lehmberg (HZSK)
Forschungsdaten in den Geisteswissenschaften (FORGE 2015), Universität Hamburg
18 September 2015, Hamburg
https://lecture2go.uni-hamburg.de/veranstaltungen/-/v/18309

A presentation of the CLARIN-D Helpdesk and the network of expertise behind it available to users of the CLARIN-D infrastructure.

4.3.13 Presentation: CLARIN Session: Morphology in CLARIN-D

Dr. Dörte de Kok, Dr. Daniël de Kok, Dr. Cagri Cöltekin
Fourth International Workshop on Systems and Frameworks for Computational Morphology
17–18 September 2015, Stuttgart
http://sfcm.eu/sfcm2015

In this session, CLARIN-D introduced the tools and services it offers. The main focus was on those tools and services directly relevant for research in the area of morphology. The presentations included general information on CLARIN-D, an overview of morphological tools and resources and a special focus on the Turkish morphology tools in WebLicht.

4.3.14 Presentation: Sprachdaten in Deutschland

Andreas Witt (IDS)
European Language Resource Coordination (ELRC) Meeting
29 September 2015, Berlin

The presentation described linguistic resources managed in IDS and discussed several questions regarding linguistic resources in general, such as which data format are typically used and where the data are stored.

4.3.15 Presentation: Recent Initiatives towards New Standards for Language Resources

Andreas Witt (IDS), Piotr Bański (IDS), Gottfried Herzog (DIN), Ulrich Heid (University of Hildesheim), Thorsten Trippel (University of Tübingen), Laurent Romary (Inria), Thomas
The presentation described three ongoing initiatives towards language resource (LR) standards coordinated and initiated by the German mirror group of ISO TC 37/SC 4 within DIN:

1. ISOTiger, an XML serialization of proposals for the syntactic annotation of text corpora,
2. “Transcription of spoken language”, a set of guidelines for transcribing spoken utterances, and
3. “Corpus Query Lingua Franca”, a meta-standard for the comparison of the formal properties of corpus query languages.

4.3.16 Presentation: Building and Annotating a Corpus of German-Language News-groups

Jasmin Schröck, Harald Lüngen (IDS)
2nd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC), Gesellschaft für Sprachtechnologie & Computerlinguistik (GSCL) 2015, Duisburg-Essen University
29 September 2015, Essen

The presentation described how to download, convert and annotate a comprehensive German news corpus from Usenet for integration in DeReKo, the German Reference Corpus hosted at the Institut für Deutsche Sprache in Mannheim.

4.3.17 Presentation: Adding Value to CMC Corpora: CLARINification and Part-of-speech Annotation of the Dortmund Chat Corpus

Michael Beißwenger (Duisburg-Essen University), Eric Ehrhardt (BBAW), Andrea Horbach (Saarland University), Harald Lüngen (IDS), Diana Steffen (Saarland University), Angelika Storrer (University of Mannheim)
2nd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC), Gesellschaft für Sprachtechnologie & Computerlinguistik (GSCL) 2015, Duisburg-Essen University
29 September 2015, Essen

The presentations described the CLARIN-D curation project “ChatCorpus2CLARIN” including the work on the integration of the Dortmund chat corpus in the CLARIN-D infrastructures, TEI representation of the resources and POS annotation using extended STTS tagset from the BMBF project “Schreibgebrauch” at Saarland University.
4.3.18 Poster: Recent Initiatives towards New Standards for Language Resources

Andreas Witt (IDS), Piotr Bański (IDS), Gottfried Herzog (DIN), Ulrich Heid (Hildesheim University), Thorsten Trippel (University of Tübingen), Laurent Romary (Inria), Thomas Schmidt (IDS), Kerstin Eckart (University of Stuttgart)

Gesellschaft für Sprachtechnologie & Computerlinguistik (GSCL) 2015, Duisburg-Essen University
1 October 2015, Essen


The poster provided an overview of three ongoing initiatives towards language resource (LR) standards coordinated and initiated by the German mirror group of ISO TC 37/SC 4 within DIN:

1. ISOTiger, an XML serialization of proposals for the syntactic annotation of text corpora,

2. “Transcription of spoken language”, a set of guidelines for transcribing spoken utterances, and

3. “Corpus Query Lingua Franca”, a meta-standard for the comparison of the formal properties of corpus query languages.

4.3.19 Presentation: Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case

Dirk Goldhahn
CLARIN Annual Conference 2015
17 October 2015, Wrocław, Poland


CLARIN offers access to digital language data for scholars in the humanities and social sciences. But how can the linguistic resources help to answer real research questions in the respective fields? By addressing research questions concerned with the work of German author Ernst Jünger an example of a successful operationalisation within the CLARIN infrastructure is introduced. In addition a new versatile Web application for answering a wide range of research questions based on vocabulary use is presented.

4.3.20 Presentation: Towards an encoding standard for social media and CMC: Experiences from German and French corpus projects using TEI

Adrien Barbaresi (BBAW), Eric Ehrhardt (University of Mannheim), Alexander Geyken (BBAW), Axel Herold (BBAW), Marc Kupietz (IDS), Lothar Lemnitzer (BBAW), Harald Lüngen (IDS), Angelika Storrer (University of Mannheim), Andreas Witt (IDS)

International Research Days: Social Media and CMC Corpora for the eHumanities
23 October 2015, Rennes, France


The contribution outlined the encoding architecture and basic features of the TEI schemas that have been designed for the CMC corpus projects ChatCorpus2CLARIN, German Wikipedia
and Usenet news corpora in DeReKo, Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK), and German WhatsApp Corpus, and gave an overview of experiences and practices in using these schemas for corpus annotation.

4.3.21 **Presentation: ‘All Your Data are Belong to Us’. European Perspective on Privacy Issues in ‘Free’ Online Machine Translation Services**

Pawel Kamocki (IDS), Jim O’Regan (Trinity College) and Marc Stauch (University of Hanover)  
**Amsterdam Privacy Conference**  
24 October 2015, Amsterdam, the Netherlands  
Presentation accepted for the Amsterdam Privacy Conference 2015, an interdisciplinary event organised by the Amsterdam Platform for Privacy Research as part of the Amsterdam Privacy Week. The paper, discussing how free on-line MT services could fit within the EU data protection framework, was presented by Pawel Kamocki during the Privacy and Technology track. It was co-authored by Jim O’Regan, an MT developer from Trinity College and Dr. Marc Stauch, an expert on privacy law from the University of Hanover. Its extended abstract as well as more information about the conference can be found at: [http://www.apc2015.net](http://www.apc2015.net)

4.3.22 **Presentation: Modeling social media and CMC genres in TEI: Models and practices from French and German corpus projects**

Thierry Chanier (Université Blaise-Pascal), Eric Ehrhardt (BBAW), Alexander Geyken (BBAW), Axel Herold (BBAW), Marc Kupietz (IDS), Lothar Lemnitzer (BBAW), Harald Lüngen (IDS), Céline Poudat (University of Nice), Angelika Storrer (University of Mannheim), Andreas Witt (IDS)  
**Connect – Animate – Innovate. Conference and Members’ Meeting of the Text Encoding Initiative (TEI), Université Lumiè re Lyon 2**  
29 October 2015, Lyon, France  
The presentation discussed how the requirements and challenges of CMC date have been handled in customized TEI schemas that have been developed for the representation of CMC and social media genres in French and German corpus projects in 2011–2015 (CoMeRe, ChatCorpus2CLARIN, German Wikipedia and Usenet news corpora in DeReKo, Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK)). The schemas developed in these projects are not independent from each other but build on each other with the goal of creating a TEI schema that fits for diverse projects in several languages.

4.3.23 **Presentation: Korpusbasierte Analyse internetbasieterer Kommunikation: Herausforderungen und Perspektiven**

Michael Beißwenger (Duisburg-Essen University), Harald Lüngen (IDS), Christian Pölitz (Technical University of Dortmund)  
**Symposium: New ways of corpora use: Data mining for text-oriented humanities, Berlin-Brandenburgische Akademie der Wissenschaften**  
30 October 2015, Berlin  
[http://www.kobra.tu-dortmund.de/mediawiki/images/2/2a/Pr%C3%A4sentation_Beisswenger_Luengen_Poelitz_30.10.2015.pdf](http://www.kobra.tu-dortmund.de/mediawiki/images/2/2a/Pr%C3%A4sentation_Beisswenger_Luengen_Poelitz_30.10.2015.pdf)
The presentation describes challenges in corpus-based analysis of computer mediated communication using data mining techniques.

### 4.3.24 Poster presentation: Aufbau von Social-Media-Korpora für die Digital Humanities: Stand und Perspektiven

Michael Beißwenger (Duisburg-Essen University), Thomas Bartz (Technical University of Dortmund), Axel Herold (BBAW), Marc Kupietz (IDS), Lothar Lemnitzer (BBAW), Harald Lüngen (IDS)

Symposium: New ways of corpora use: Data mining for text-oriented humanities, Berlin-Brandenburgische Akademie der Wissenschaften
30 October 2015, Berlin

[http://www.kobra.tu-dortmund.de/mediawiki/images/7/7e/Poster_Beisswenger_et_al_Social_Media_Corpora_30.10.2015.pdf](http://www.kobra.tu-dortmund.de/mediawiki/images/7/7e/Poster_Beisswenger_et_al_Social_Media_Corpora_30.10.2015.pdf)

The poster illustrates the needs of standards for building social media corpora for digital humanities. It introduces a TEI scheme for computer mediated communication corpora and an NLP method for linguistic annotation of social media genre using extended POS STTS 2.0 tagset.

### 4.3.25 Presentation: Vorstellung des Langzeitvorhabens gefördert im Akademienprogramm: “Grammatiken, Korpora und Sprachtechnologie für indigene nordeurasische Sprachen”

Beáta Wagner-Nagy (IFUU/UHH), Timm Lehmberg, Hanna Hedeland (HZSK)

Forschungskolloquium “eHumanities”, Universität Hamburg
17 December 2015, Hamburg

A presentation of the project INEL (Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages) and its integration into the HZSK and CLARIN-D infrastructure.

### 4.3.26 Presentation: CLARIN webservice (Übung)

Nils Reiter, Sarah Schulz (IMS, Uni Stuttgart)

Course Computerlinguistische Methoden für Digital Humanities
Winter term 2015/2016, Stuttgart

Presentation and exercise using CLARIN web services for text processing. Students used segmentation tools and part of speech taggers to extract features on literary texts.

### 4.3.27 Presentation: Wikipedia Resources at IDS Mannheim for Linguistic Research: Corpora, Tools, Annotations

Harald Lüngen, Eliza Margaretha (IDS)

Wikipedia: Discourse and corpus linguistic perspectives Workshop
20 January 2016, Mannheim

The presentation discussed the TEI customization and CMC features in the Wikipedia article and discussions corpora in IDS. Besides, it explained the conversion pipeline and posting segmentation used to build the corpora in I5 format from Wikipedia dumps in Wikitext format.
4.3.28 Lecture: CLARIN: Research Infrastructure for the Humanities and Social Sciences

Prof. Dr. Erhard Hinrichs, Dr. Thorsten Trippel
High Precision Philology in a Digital Space
15–16 February 2016, Leipzig
In this lecture, CLARIN-D introduced the tools and services it offers. The main focus was on those tools and services directly relevant for classic research.

4.3.29 Poster presentation: Kuration und Exploration des Korpus “Diskurs in der Weimarer Republik”

Peter Fankhauser (IDS)
Digital Humanities im Deutschsprachigen Raum (DHD) 2016
9 March 2016, Leipzig
http://www.dhd2016.de/abstracts/posters-056.html
This poster described the curation of the corpus “Discourse in the Weimar Republic” according to CLARIN-guidelines, and demonstrated how the resulting curated corpus can be used for explorative contrastive analysis.

4.3.30 Poster presentation: Lizenzauswahlwerkzeuge für die digitalen Geisteswissenschaften (License Selectors for Digital Humanities)

Pawel Kamocki (IDS), Andreas Witt, Erik Ketzan (IDS)
DHD 2016 (Die 3. Tagung des Verbands Digital Humanities im deutschsprachigen Raum)
9 March 2016, Leipzig
A poster about various License Selectors (applications that help the user choose a license) and its usefulness for Digital Humanities.

4.3.31 Poster Presentation: Das Dortmunder Chat-Korpus in CLARIN-D: Modellierung und Mehrwerte

Michael Beißwenger (Duisburg-Essen University), Axel, Herold (BBAW), Harald, Lüngen (IDS), Angelika Storrer (University of Mannheim)
Digital Humanities im Deutschsprachigen Raum (DHD 2016)
9 March 2016, Leipzig
The poster illustrates how the Dortmund chat corpus is modelled for the integration in the CLARIN-D infrastructure including archiving the resources in the BBAW and IDS repositories, registering the metadata in Virtual Language Observatory (VLO) and providing access to the corpus data via Federated Content Search (FCS). It also explains how the corpus is extended for interoperability in digital humanities, for instance with the existing spoken and written language corpora in BBAW and IDS.

4.3.32 Presentation: Operationalisierung von Forschungsfragen in CLARIN-D – Der Anwendungsfall Ernst Jünger

Dirk Goldhahn
CLARIN offers access to digital language data for scholars in the humanities and social sciences. But how can the linguistic resources help to answer real research questions in the respective fields? By addressing research questions concerned with the work of German author Ernst Jünger an example of a successful operationalisation within the CLARIN infrastructure is introduced. In addition a new versatile Web application for answering a wide range of research questions based on vocabulary use is presented.

4.3.33 Presentation: A Frankenstein of quotations—is it alive? To what extent can the quotation exception be used to build language corpora under EU law?

Pawel Kamocki (IDS)
LREC 2016, Legal Issues Workshop
24 May 2016, Portoroz (Slovenia)
A presentation about how the quotation exception (i.e. a copyright exception that allows quotations) can be used to build language corpora under EU law. The presentation takes into account recent developments such as the Painer case (CJEU case C-145/10) and the copyright reform in Slovakia.

4.3.34 Poster presentation: KorAP Architecture—Diving in the Deep Sea of Corpus Data

Nils Diewald, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Banski and Andreas Witt (IDS)
LREC Main Conference 2016
25–27 May 2016, Portoroz (Slovenia)
This poster illustrates the architecture of KorAP system which is designed to be scalable, flexible, and sustainable to handle very large corpora with multiple annotations, particularly the German Reference Corpus DeReKo.

4.4 Working with Corpora (WwC)

Working with Corpora is a pilot experience of internal training on corpus linguistics methodologies at the University of Saarland. In-house researchers run free training in reproducible methods creating a cooperative and collaborative dynamic. Courses are self-managed by participants and they are open to both researchers and students.

The focus is on practical matters (the nitty-gritty stuff that nobody mentions in papers but which takes a lot of time) around corpus-based work, language technology, etc. ranging from conceptual topics like corpus design, pipeline design and good practices, to concrete topics (XML, CQP, WebLicht...).

The goals of the training are:

- rising awareness about tools and methods available,
- empowering colleagues to carry out corpus-based research/teaching,
• scaffolding their progress, specially at the beginning

More information about Working with Corpora at UdS can be found on the website. There have been two series of WwC sessions:

• SS 2015
• WS 2015/2016

For the WS 2015/2016, we used DIT4C cloud technology deployed at ResBaz so that nobody had to worry about installation/operating system/specs problems. All code is written and executed within Jupyter Notebook allowing easy access to earlier input and output, as well as the rich display of text/images.

Materials have been added to TeLeMaCo.

4.4.1 WwC SS2015: Text preprocessing (English and German) with WebLicht

José Manuel Martínez Martínez
WwC: Text preprocessing (English and German) with WebLicht
18 May 2015, Universität des Saarlandes
http://fedora.clarin-d.uni-saarland.de/unserwiki/doku.php?id=training:weblicht

• What is WebLicht? Web services, and main concepts
• Getting started
• Input selection
• Tools, tool chains and modes
• Examination of results
• Exercises
• Output format
• Exploring results with TüNDRA
• Manual edition of linguistic annotation with WebAnno
• Advanced WebLicht (command line)

4.4.2 WwC SS2015: Querying corpora with CQP/CQPweb

José Manuel Martínez Martínez
WwC: Querying corpora with CQP/CQPweb
6 July 2015, Universität des Saarlandes
http://fedora.clarin-d.uni-saarland.de/unserwiki/doku.php?id=training:cqp

• What is CQP?
• What is CWB?
• What is CQPweb?
• Queries:
  – word forms, POS, lemma
  – multi word queries
  – regular expressions
  – conditions (AND, OR)
  – filtering by metadata
• CQPweb facilities: frequencies, distributions, collocations, sorting
• Saving results and basic post-processing

4.4.3 WwC SS2015: Analysing and visualising data with R I
José Manuel Martínez Martínez
13 July 2015, Universität des Saarlandes
http://fedora.clarin-d.uni-saarland.de/unserwiki/doku.php?id=training:analyzing_and_visualising_data

  • Getting your data in R
  • Preparing your data in R
  • Describing your data in R
  • Basic visualization of data in R: boxplots, barplots, line plots
  • Hypothesis testing

4.4.4 WwC SS2015: Analysing and visualising data with R II
Ekaterina Lapshinova-Koltunski
31 July 2015, Universität des Saarlandes
http://fedora.clarin-d.uni-saarland.de/unserwiki/doku.php?id=training:analyzing_and_visualising_data_2

  • Getting your data in R
  • Preparing your data in R
  • Clustering techniques
    – how to do it
    – how to interpret the results
  • CA techniques
    – how to do it
    – how to interpret the results
4.4.5 WwC SS2015: Reporting statistics in different output formats with R

José Manuel Martínez Martínez
17 July 2015, Universität des Saarlandes
http://fedora.clarin-d.uni-saarland.de/unserwiki/doku.php?id=training:cross_format_stats_report

- Saving data frames as plain text, CSV, Excel, Word, PowerPoint.
- Saving plots as PDF, images, as editable vectorial graphics, as Word, as PowerPoint.
- What is R markdown?
- Producing a R markdown version and getting different outputs: HTML, PDF, Word, LaTeX, slides, markdown

4.4.6 WwC WS2015/2016: Data Carpentry with Python and NLTK

Daniel McDonald, José Manuel Martínez Martínez, Liling Tan
6 November, 20 November, 27 November, 4 December, 11 December, 18 December, 15 January, 22 January, 5 February, 12 February
WwC: Data Carpentry with Python and NLTK
https://github.com/interrogator/wwc

- Session 1: orientation, Python, Jupyter notebooks, NLTK
- Session 2: sentence splitting, tokenisation, keywords, n-grams, collocates, concordancing
- Session 3: PoS tagging, lemmatisation, exploring annotated data, writing a concordancer
- Session 4: regular expressions
- Session 5: HTML and XML
- Session 6: file operations
- Session 7: using Git

4.4.7 SaCoCo technical tutorial

José Manuel Martínez Martínez
http://chozelinek.github.io/sacoco/
Technical tutorial showing step-by-step how to use the CLARIN-D infrastructure to compile a diachronic corpus of German cooking recipes.

- data acquisition and preprocessing
- linguistic annotation with WebLicht
- corpus encoding for CQPweb
- research question formulation
- operationalisation of the research question
- feature extraction with CQPweb
- visualization and analysis of the data with CQPweb/R.
5 TeLeMaCo

TeLeMaCo is a collaborative portal for all kinds of training and teaching materials relevant to linguistics and digital humanities, which was described in more detail in the annual reports R8.2 and R8.4.

In the last year we saw a steady growth of materials described in TeLeMaCo. We implemented access monitoring and see approximately 300 accesses per month.

We presented TeLeMaCo at various events including the Forschungsdaten in den Geisteswissenschaften (FORGE 2015), the European summer university Digital Humanities Culture & Technology and the Fall school in Tbilisi (Georgia). A new kind of material featured in TeLeMaCo are detailed technical tutorials, which lead the reader step-by-step through a particular process including all necessary commands and scripts. These tutorials may be used in class, for self-study or for documenting reproducible research. Examples are the materials developed during the Working with corpora workshop Data Carpentry with Python and NLTK and the SaCoCo technical tutorial (see descriptions above).