ChatCorpus2CLARIN: Project background
Curation project of the CLARIN-D F-AG 1 “German Philology”

**Duration:** May 2015 – February 2016

**The task:** develop a workflow and resources for the integration of an existing chat corpus into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences ([http://clarin-d.de](http://clarin-d.de)).

**Project team:** Michael Beißwenger (U Dortmund / DUE), Angelika Storrer, Eric Ehrhardt (U Mannheim), Harald Lüngen (IDS Mannheim), Axel Herold (BBAW, Berlin) + other colleagues at the CLARIN-D hubs at IDS and BBAW.

[http://www.clarin-d.de/en/curation-project-1-3-german-philology](http://www.clarin-d.de/en/curation-project-1-3-german-philology)
Common Language Resources and Technology Infrastructure

Vision: all digital language resources and tools from all over Europe and beyond should be accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences

Backbone: federation of language data repositories, service and knowledge centers

Important features:

• Interoperability of resources and tools across centers
• Services for the integration of resources and tools
• Development and provision of state-of-the-art tools for annotation and exploitation of resources

19 European member countries
ChatCorpus2CLARIN: Project background
Curation project of the CLARIN-D F-AG 1 “German Philology”

**Duration:** May 2015 – February 2016

**The task:** develop a workflow and resources for the integration of an existing chat corpus into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences (http://clarin-d.de).

**Project team:** Michael Beißwenger (U Dortmund / DUE), Angelika Storrer, Eric Ehrhardt (U Mannheim), Harald Lüngen (IDS Mannheim), Axel Herold (BBAW, Berlin) + other colleagues at the CLARIN-D hubs at IDS and BBAW.

http://www.clarin-d.de/en/curation-project-1-3-german-philology
Dortmund Chat Corpus  http://www.chatkorpus.tu-dortmund.de

- **scope:** Language use and linguistic variation in German chats
- **corpus size:** 478 logfiles with 140240 user posts / 1 million words
- **collection maxim:** build a corpus which demonstrates the bandwith of usage contexts of chat technology ("It’s not the technological constraints (alone) that determine a certain use of the German language – it’s the parameters of the social use of the technology!")

- **content:** chat discourse from heterogeneous sources representing the use of chats in a wide range of application contexts (social chats, advisory chats, chats in the context of learning and teaching, chats in the media context)

- **availability:** online for download since 2005 (together with a simple query tool, addressees: linguists) + as a collection of HTML pages (for online browsing, addressees: German teachers)
Dortmund Chat Corpus

- **scope:** Language use and linguistic variation in German chats
- **corpus size:** 478 logfiles with 140,240 user posts / 1 million words
- **collection maxim:** Build a corpus which demonstrates the bandwidth of usage contexts of chat technology ("It's not the technological constraints (alone) that determine a certain use of the German language - it's the parameters of the social use of the technology!")
- **content:** Chat discourse from heterogeneous sources representing the use of chats in a wide range of application contexts (social chats, advisory chats, chats in the context of learning and teaching, chats in the media context)
- **availability:** Online for download since 2005 (together with a simple query tool, addressees: linguists) + as a collection of HTML pages (for online browsing, addressees: German teachers)
The original resource (chat corpus 1.0)

XML-annotated on basis of a homegrown XML format (‘ChatXML’) which describes:

1. the basic structure and properties of logfiles and postings (“messages”)
2. selected items on the micro-level of user posts (emoticons, acronyms, addressing terms, nickname mentions)
3. selected metadata about the chat platforms and users.
Goals of the project

Primary goals:

- **Sustainability**: Integrate the only CMC corpus for German which is freely available (and which is used by many researchers) into the CLARIN-D corpus infrastructure

- **Enhanced query options**: Add value through additional linguistic annotations (part-of-speech)

- **Interoperability**: Represent the corpus compliant to an established standard in the Digital Humanities and thus make it interoperable with other resources in the CLARIN-D corpus infrastructure

  ⇒ **Advanced options for comparative analyses** of CMC with other types of discourse documented in other types of corpora (text and speech corpora)
Goals of the project

Secondary goals:

- Create a showcase which demonstrates what researchers can gain when
  - CMC corpora are made available for the community,
  - CMC corpora – as part of big, annotated corpus collections – can be analyzed in combination with other language resources (text and speech corpora at IDS and BBAW)

- Intended model character of the solutions developed in the project: solutions should be useful not only for the modeling and integration of the chat corpus but also for the modeling and integration of other CMC corpora into CLARIN-D (future projects)
Main requirements to reach the goals

- **interoperability and sustainability:**
  - represent the corpus according to established standards in the Digital Humanities
  - TEI conversion (using schemas from the TEI CMC-SIG)

- **integration into the corpus infrastructures of CLARIN-D institutions:** clear the legal constraints for republishing the resource in CLARIN-D
  - Legal opinion (John H. Weitzmann / iRights Law)

- **part-of-speech tagging:** adapt NLP tools for CMC
  - Extended STTS tag set (using a CMC-adapted tag set from the EmpiriST shared task and a tool chain from the project www.schreibgebrauch.de)
Spotlight 1: TEI representation

Why TEI? — we really need interoperability of resources through standardisation!

- TEI is the de facto standard for text encoding in Digital Humanities (first version 1990)
- TEI offers a large and very lively international community …and tools
- TEI is also used for encoding linguistic corpora, including the corpora in CLARIN-D
- and including the corpus holdings at both project partners BBAW and IDS
The official TEI lacks models for the representation of the concepts and structural features of CMC.

The TEI SIG CMC (since 2013) seeks to close this gap.

Two *TEI customisations* for CMC corpus projects had been developed by members of the TEI SIG CMC:
- DeRiK schema (2012, Germany)
- CoMeRe schema (2014, France)
CLARIN-D TEI schema for CMC

Developed project-specific TEI customisation, called **CLARIN-D CMC-TEI**, based on

- CoMeRe schema (= most recent schema from TEI SIG CMC)
- Analysis of logfiles in DO chat corpus
- Analysis of samples of other CMC genres: Twitter, WhatsApp, Usenet news, and Wikipedia talk pages

Documentation of the schema in the TEI Wiki:
CLARIN-D TEI schema for CMC

□ Customisation_1: Introduction of specific models for CMC-specific concepts: e.g. <post>, @auto

□ Customisation_2: Modification of standard models which are used more flexible in CMC than in traditional written genres (e.g. <opener> <closer>, <postscript>)

□ Definition of best practices for the use of standard TEI models (without modification; e.g. <div>, <w>, @type, <participantList>, <timeline>
who="#f1101004.A01_System" xml:id="f1101004.m137">
<time>22:01</time>
<name corresp="#f1101004.A04" type="nickname">[_MALE-TEACHER-A04_]</name>
entered the room <name type="roomname">[_ROOMNAME_]</name> at 22:01:55
</post>

<post auto="false" rend="color:black" synch="#f1101004.t047" type="standard"
who="#f1101004.A03" xml:id="f1101004.m138">
<time>22:02</time>
<anchor type="sentence_start"/>
<w lemma="gut" type="ADJD" xml:id="f1101004.m138.t1">gut</w>
<w lemma="," type="$," xml:id="f1101004.m138.t2">,</w>
<w lemma="die" type="PDS" xml:id="f1101004.m138.t3">das</w>
<w lemma="sollen" type="VMFIN" xml:id="f1101004.m138.t4">sollte</w>
<w lemma="unser" type="PPOSAT" xml:id="f1101004.m138.t5">unseren</w>
<w normal="Arbeitseifer" type="NN" xml:id="f1101004.m138.t6">arbeitseifer</w>
<w lemma="nicht" type="PTKNEG" xml:id="f1101004.m138.t7">nicht</w>
<w lemma="stören" type="VVINF" xml:id="f1101004.m138.t8">stören</w>
[...]
</post>
22:01

[__MALE-TEACHER-A04__] entered the room at 22:01:55

22:02

[gut],
das sollen unseren Arbeitseifer nicht stören]
Spotlight 2: Legal expertise -- Mandate

- Sought legal opinion on republishing the resource in CLARIN-D
- Law firm *iRights law* (lawyer John H. Weitzmann)
- Mandate: check legal restrictions arising from IPRs, copyright, personality rights and other legal statuses
- Material: Detailed documentation of all subcorpora, including description of the original chat platforms and collection procedures; corpus samples
Spotlight 2: Legal expertise -- Results

- No concerns regarding copyright – the overwhelming majority of chat posts do not constitute works of art.
- Chat participants and mentioned individuals (with the exception of public figures) must not be identified → anonymise names, nicknames, geographical names, host names, IP addresses etc.
- Subcorpus of 8 logfiles of psycho-social counselling chats to be removed altogether.
- EU law on protection of databases applies to the collection, preparation, and curation of the DO chat corpus 1.0 and 2.0 → provide resource with license CC BY 4.0.
Spotlight 2 -- Anonymisation

- Automatically replace names with category labels using reference to participantList in teiHeader if available

```
<name corresp="#f1101006.A08" type="nickname">
  <w lemma="Frau" type="NN" xml:id="m234.t6">Frau</w>
  <w type="NE" xml:id="m234.t7">[_FEMALE-TEACHER-A08_]</w>
</name>
```

- Separate manual campaign: 4 student assistants mark remaining sensitive names and references and provide them with a category label using the author mode in Oxygen XML editor
Result: CLARIN-D-conformant resource

<table>
<thead>
<tr>
<th>Dortmund Chat corpus 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td># chat log files</td>
</tr>
<tr>
<td># posts</td>
</tr>
<tr>
<td># tokens</td>
</tr>
<tr>
<td>file Size (TEI-XML)</td>
</tr>
</tbody>
</table>
Availability of the integrated resource

- Integration in CLARIN-D repository at IDS: done
- Integration in CLARIN-D repository at BBAW: this week
- Downloadable only when full anonymisation is finished
Availability of the integrated resource

- To be integrated in the German Reference Corpus DEReKo at IDS and searchable through COSMAS II
- To be integrated in the DWDS corpus query interface at BBAW
- Will also be searchable via CLARIN web services and the CLARIN Federated Content Search
Outlook and future work

TEI representation

- The corpus can be used as a showcase for the representation of a CMC corpus with the CLARIN-D TEI schema for CMC
- The schema can be used as an input for the further discussion process in the TEI CMC-SIG on what a TEI standard for CMC should look like

Legal opinion and anonymisation

- Need *automated* anonymisation which is legally safe

CMC Corpora

- Integration of more CMC resources in the CLARIN-D research infrastructures
- Remodeling of resources already integrated (Wikipedia talk corpus, Usenet corpus) using the CLARIN-D CMC-TEI
(Best) Practices for Annotating and Representing CMC and Social Media Corpora in CLARIN-D

Thank you!

cmccorpora16:
4th Conference on CMC and Social Media Corpora for the Humanities
University of Ljubljana
September 27—28, 2016
ChatXML to TEI conversion

Set of XSLT stylesheets

Quality assurance

- Log file of the conversion
- Primary data diff
Outlook and future work

Legal issues and anonymization:

- Problem: Linguists are laymen when it comes to the assessment of legal issues regarding the collection and republishing of CMC data.

- Desideratum: A general legal opinion commissioned and disseminated by and via an acknowledged language resources initiative would be an important prerequisite for the further development of the CMC corpora landscape and community (We will communicate this issue at the CLARIN Conference in Aix-en-Provence, this October!)

Future work:

- Integration of further CMC resources into the CLARIN-D corpus infrastructures.

- Remodeling of resources already integrated (Wikipedia talk corpus, Usenet corpus) using the CLARIN-D TEI schema.)