

# (Best) Practices for Annotating and Representing CMC and Social Media Corpora in **CLARIN-D**

Michael Beißwenger, Eric Ehrhardt,  
Axel Herold, Harald Lüngen,  
Angelika Storrer

UNIVERSITÄT  
DUISBURG  
ESSEN

*Offen im Denken*

UNIVERSITÄT  
MANNHEIM



INSTITUT FÜR  
DEUTSCHE SPRACHE



berlin-brandenburgische  
AKADEMIE DER WISSENSCHAFTEN

**cmccorpora16:**  
**4th Conference on CMC and Social Media Corpora  
for the Humanities**

University of Ljubljana

September 27—28, 2016

# 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 Längen (IDS Mannheim), Axel Herold (BBAW, Berlin) + other colleagues at the CLARIN-D hubs at IDS and BBAW.

The screenshot shows the CLARIN-D website interface. At the top, there are flags for Germany and the UK, followed by the CLARIN-D logo. To the right, there are navigation icons for Accessing, Analysing, Preparation, More, and Help. Below the navigation bar, the page title is "ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D" and the sub-heading is "Project content". The main text describes the project's goal: to restructure an existing corpus of computer-mediated communication (CMC) to conform to current standards for representation in the Digital Humanities context. The project will (1) transform the metadata and annotations of the chat corpus into a TEI-compliant format, (2) enrich the data with further linguistic annotations, and (3) integrate the resulting resource into the CLARIN-D Corpus Infrastructures at the Institute for the German Language (IDS) and the Berlin-Brandenburg Academy of Sciences (BBAW). The footer text states that the integration in CLARIN-D will allow for a systematic corpus-based analysis of CMC discourse as compared to the language of edited text (as represented in the text corpora at BBAW and IDS) and of spoken conversations (as represented in the spoken language corpora at IDS).

<http://www.clarin-d.de/en/curation-project-1-3-german-philology>

# About CLARIN

*clarin.eu*

## 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 Längen (IDS Mannheim), Axel Herold (BBAW, Berlin) + other colleagues at the CLARIN-D hubs at IDS and BBAW.

The screenshot shows the CLARIN-D website interface. At the top, there are flags for Germany and the UK, followed by the CLARIN-D logo. A navigation bar includes icons for Accessing, Analysing, Preparation, More, and Help. Below this is a dark red navigation menu with links for Home, Accessing, Analysing, Preparation, Disciplines, About, and Help. The main content area features the title "ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D" and a sub-heading "Project content". The text describes the project's goal to integrate chat corpora into the CLARIN-D infrastructure for linguistic analysis. It mentions the transformation of metadata and annotations, enrichment with linguistic annotations, and integration into the CLARIN-D Corpus Infrastructures at the Institute for the German Language (IDS) and the Berlin-Brandenburg Academy of Sciences (BBAW). A footer note states that the integration will allow for a systematic corpus-based analysis of CMC discourse compared to edited text.

<http://www.clarin-d.de/en/curation-project-1-3-german-philology>

# The original resource (chat corpus 1.0)

**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 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)

## Dortmund Chat Co

- **scope:** Language u
- **corpus size:** 478 lo
- **collection maxim:** l  
bandwith of usage c  
technological constr  
the German languag  
technology!")
- **content:** chat disco  
the use of chats in a  
chats, advisory chat  
teaching, chats in th
- **availability:** online f  
simple query tool, ac  
pages (for online browsing, addressees: *German teachers*)



## Dortmunder Chat-Korpus

Bestand

Korpora / Download

Recherche

MIT STACCA<sup>Do</sup>

Kontakt

Das **Dortmunder Chat-Korpus** dokumentiert anhand einer Sammlung von Mitschnitten (sog. "Logfiles") die Sprachverwendung in unterschiedlichen Typen von Chat-Anwendungen. Es ist als Grundlage und Hilfsmittel für sprachwissenschaftliche Untersuchungen zur synchronen internetbasierten Kommunikation konzipiert und wird in verschiedenen Versionen zur freien Nutzung zur Verfügung gestellt.

Das Korpus umfasst mit über 140.000 Chat-Beiträgen bzw. 1,06 Millionen laufenden Wortformen umfangreiches Datenmaterial aus diversen Einsatzformen der Chat-Technologie. Der Bestand reicht von **Chats im Hochschulkontext** (E-Learning, Online-Zusammenarbeit, kollektive Experten-Interviews) und im Praxisbereich **Beratung & Support** über **Chat-Events im Medienkontext** (Chats mit Politikern und Medienakteuren oder begleitend zu TV-Ereignissen) bis hin zu **"Plauder"-Chats im Freizeitbereich**, die im IRC-Netzwerk oder in **Webchat-Communities** stattgefunden haben. Die Korpusdokumente wurden anhand einer XML-Sprache für Recherchezwecke aufbereitet.

Zusammen mit dem Korpus wird ein Suchwerkzeug zur Verfügung gestellt: **STACCA<sup>Do</sup>** ermöglicht es, auf einfache Weise nach chat-typischen Elementen wie z.B. Emoticons, Adressierungen, Asterisk-Ausdrücken oder Zuschreibungen ("action messages") zu recherchieren, beliebige einfache und komplexe Volltext-Suchanfragen zu formulieren oder statistische Auswertungen zum Kommunikationsaufkommen und zum Beitragsverhalten einzelner Chatter in den Teilkorpora oder in einzelnen Korpusdokumenten zu erzeugen.

**Wenn Sie unsere Website zum ersten Mal besuchen** und einfach nur mal in unserem Datenbestand stöbern möchten, können Sie auf 385 Dokumente aus unserem Korpus auch bequem per Browser zugreifen: [HTML-Version des Releasekorpus](#)

Das **Dortmunder Chat-Korpus** ist Ergebnis eines Lehrstuhlprojekts am Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik, das unter der Leitung von Prof. Dr. Angelika Storrer und Dr. Michael Beißwenger am Institut für deutsche Sprache und Literatur der Technischen Universität Dortmund realisiert wurde. Das Suchwerkzeug STACCA<sup>Do</sup> wurde von Bianca Stockrahm programmiert.

**Kurzbeschreibungen des Dortmunder Chat-Korpus finden sich in den folgenden Publikationen:**

- Beißwenger, Michael; Storrer, Angelika (2008): **Corpora of Computer-Mediated Communication**. In: Anke Lüdeling & Merja Kytö (Eds): *Corpus Linguistics. An International Handbook*. Volume 1. Berlin. New York (Handbooks of Linguistics and Communication Science 29.1), 292-308.
- Beißwenger, Michael; Storrer, Angelika (2011): [Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen - Erfahrungen - Desiderate](#). In: *Journal for Language Technology and Computational Linguistics*, 119-139.
- Beißwenger, Michael (2013): **Das Dortmunder Chat-Korpus**. In: *Zeitschrift für germanistische Linguistik* 41/1, 161-164.

# The original resource (chat corpus 1.0)

- ```
- <message color="seagreen" creator="Andra" type="utterance" id="359">
  <timestamp> 20:25:32 </timestamp>
  - <messageHead>
    <nickname>Andra</nickname>
  </messageHead>
  - <messageBody>
    mein traum ist ja zugegeben: einmal in Chile degus in freier natur beobachten, das wäre genial!
    
  </messageBody>
</message>
- <message color="green" creator="Denise"
  <timestamp> 20:25:33 </timestamp>
  - <messageHead>
    <nickname>Denise</nickname>
  </messageHead>
  - <messageBody>
    hallöchen
    <nickname>Melanie</nickname>
  </messageBody>
</message>
- <message color="navy" creator="Jasmin"
  <timestamp> 20:25:46 </timestamp>
  - <messageHead>
    <nickname>Jasmin</nickname>
  </messageHead>
  - <messageBody>
    <address addressee="baloo">@bal
    schön zu hören
  </messageBody>
</message>
```
- 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](http://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

# Spotlight 1: TEI representation

- 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:

<http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema>

# 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>)

# TEI representation

```
<post auto="true" rend="color:blue" synch="#f1101004.t046" type="event"
  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>
```

# TEI representation

```
<post auto="true" end="color:blue" synch="#f1101004.t046" type="event"
  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" end="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>
```



## 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

Dortmund Chat corpus 2.0	
# chat log files	470
# posts	131,033
# tokens	1,005,166
file Size (TEI-XML)	100MB

# 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

The image shows two screenshots. The left one is a browser window displaying a Fedora repository entry for 'Dortmunder Chatkorpus 2.0'. The right one is a sign-in interface for the CLARIN Service Provider Federation.

**Browser Window: Dortmund Chatkorpus 2.0**

URL: [repos.ids-mannheim.de/fedora/objects/clarin-ids:chat.000000/datastreams/CMDI/content](http://repos.ids-mannheim.de/fedora/objects/clarin-ids:chat.000000/datastreams/CMDI/content)

**Dortmunder Chatkorpus 2.0**

[de] Diese Ressource umfasst das Dortmunder Chatkorpus 2.0: 470 Chat Logfiles aus den Jahren 2000-2006 mit 1.005.166 Tokens in 131.033 Chat-Nachrichten aus soziale Chats, Beratungschats, Chats im Kontext von Universitätsseminaren und moderierte Chats im Medienkontext. Strukturelle Annotation nach CLARIN CMC-TEI; POS-Annotation nach STTS 2.0

[en] This resource comprises the Dortmund Chat Corpus 2.0: 470 chat logfiles from the years 2000-2006 containing 1,005,166 tokens in 131,033 chat messages from social chat, advisory chat, chat in university teaching and moderated chats related to media. Structural annotation according to CLARIN CMC-TEI; POS annotation according to STTS 2.0.

**Metadata:**

- PID: <http://hdl.handle.net/10932/00-033B-0995-19D0-0A01-B>
- Type: collection
- Issued: 2016-07-29
- Creator: CLARIN-D curation project ChatCorpus2CLARIN: Michael Beißwenger, Angelika Storrer, Eric Ehrhardt, Axel Herold, Harald Längen
- Publisher: Institut für Deutsche Sprache und Berlin-Brandenburgische Akademie der Wissenschaften

**Relations:**

- LandingPage: <http://www.clarin-d.de/de/kurationsprojekt-1-3-germanistik>
- relation: Harald Längen, Michael Beißwenger, Eric Ehrhardt, Axel Herold, Angelika Storrer (To appear in September 2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In Proceedings of KONVENS 2016. Ruhr-Universität Bochum.
- hasPart: Corpus (tgr, 12.8 MB, restricted)

**Metadata Information**

- Creator: HL
- Creation date: 2016-08-12
- Profile: [clarin.eu/cr1.p\\_13668895758244](http://clarin.eu/cr1.p_13668895758244)
- Collection: Institut für Deutsche Sprache, CLARIN-D Zentrum, Mannheim

**Sign-in via the CLARIN Service Provider Federation**

Select your identity provider below. This is usually the institution where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in.

If you cannot find your institution in the list below, please select the *clarin.eu* website account and use your CLARIN website credentials. If you don't have such credentials you can register an account [here](#). For questions please contact [spi@clarin.eu](mailto:spi@clarin.eu).

**Identity Providers:**

- University of Ljubljana (Slovens)
- University of Ljubljana, Faculty of Medicine (Slovens)

**Search:** Ljub

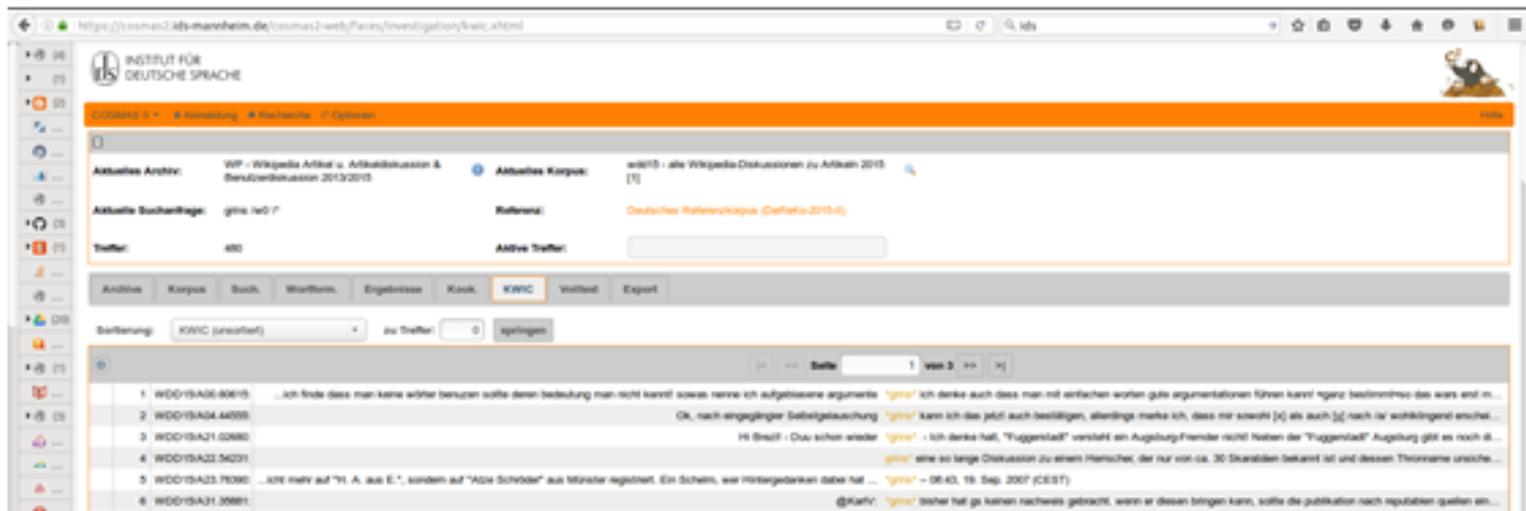
Locate me and show nearby providers

Show providers in: all countries

Based on Disclosure © UNINETT  
Version: 1.9.1, Server ID: 342

# 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



The screenshot displays the COSMAS II web interface for the German Reference Corpus. The browser address bar shows the URL: <https://cosmas2.ids-mannheim.de/cosmas2-web/Faces/investigation/kwic.xhtml>. The page header includes the logo of the Institut für Deutsche Sprache and the text "COSMAS II".

The search parameters are as follows:

- Actual search term: ginn /gO7/
- Actual corpus: wR15 - alle Wiggels Diskussionen zu Artfakt 2015
- Reference: Deutsches Referenzkorpus (DerReKo-2015-6)
- Truffer: 480
- Active Truffer: (empty)

The search results are displayed in a table with the following columns: ID, Corpus, and Text. The results are filtered by the keyword "KWIC".

ID	Corpus	Text
1	WDD-ISA06.0019	...ich finde dass man keine wörter benutzen sollte deren bedeutung man nicht kennt! sowas nennt ich aufgelassene argumente "ginn" ich denke auch dass man mit einfachen wörtern gute argumentationen führen kann! nganz bestimmt die was ernt m...
2	WDD-ISA04.4038	Oh, nach eingängiger Selbstbezeichnung "ginn" kann ich das jetzt auch bestätigen, allerdings merke ich, dass mir sowohl [x] als auch [y] nach 'a' wichtig sind ansteh...
3	WDD-ISA21.0280	Hilf! Du schon wieder "ginn" - ich denke halt, "Fuggenlaß" versteht ein Augsburg-Fremde nicht! Neben der "Fuggenlaß" Augsburg gibt es noch B...
4	WDD-ISA22.5621	"ginn" eine so lange Diskussion zu einem Fremder, der nur vor ca. 30 Sekunden bekannt ist und dessen Thematik unklar ist...
5	WDD-ISA23.7030	...ist neiv auf "H. A. aus E.", sondern auf "Ade Schöder" aus Münster registriert. Ein Schein, wie Mitsagedanken dabei hat ... "ginn" - 05.10.19. Sep. 2007 (CEST)
6	WDD-ISA31.3081	@Kerli/ "ginn" bisher hat ja keinen nachweis gebracht, wenn er diesen bringen kann, sollte die publikation nach reputieren gehen en...

# 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!***

---

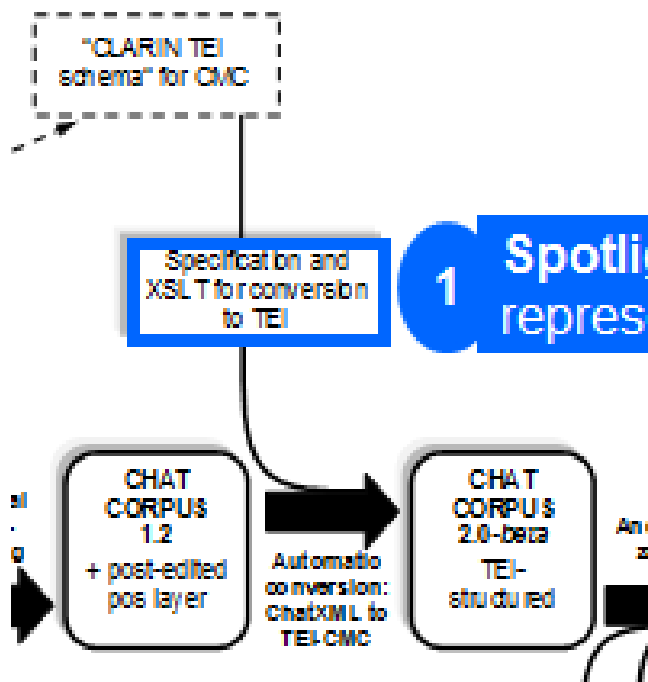
**cmccorpora16:  
4th Conference on CMC and Social Media Corpora  
for the Humanities**

University of Ljubljana

September 27—28, 2016



# ChatXML to TEI conversion



## Quality assurance

- Log file of the conversion
- Primary data diff

Set of XSLT stylesheets

# 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)