Schemas and experiences from modeling German CMC empirical Corpora in TEI

Chat-Korpus
www.chatkorpus.tu-dortmund.de

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



in connection with

Adrien Barbaresi, Alexander Geyken, Marc Kupietz, Lothar Lemnitzer, Andreas Witt



International Research Days: Social Media and CMC Corpora for the eHumanities



Journées Internationales de recherche «Médias sociaux et corpus de communication médiée par les réseaux. Annotation, analyse, données libres» 23-24 octobre 2015

TEI schema drafts of the SIG: Some basic features

DeRik schema (Beißwenger, Ermakova, Geyken, Lemnitzer, Storrer 2012):

- Introduction of an element model <post> for written user contributions to CMC interactions which combines features of text divisions and spoken utterances.
- Adaptation of the existing element model <div> for the rerpesentation of CMC threads and logfiles.
- Introduction of diverse models for CMC phenomena below the <post> level.

CoMeRe schema (Chanier, Poudat, Sagot, Antoniadis, Wigham 2014):

- Introduction of an element prod> for the representation of non-verbal acts
- (re-)efinition of <post>, <prod> and <u> as models which may be combined within one interaction. ⇒ make the schema fit for multimodal CMC

CLARIN-D schema (Beißwenger, Ehrhardt, Herold, Lüngen, Storrer 2015):

- Reduction of new elements through re-modeling some CMC-specific concepts from the previous schemas with "standard" TEI
 - ⇒ guiding principle: "reduce to the max": introduction of new models and modification of existing models only for concepts which are needed *in any case*; for everything else: best practices for the use of existing models
- Definition of an interface to part-of-speech annotations

CLARIN-D curation project «ChatCorpus2CLARIN»





Home Accessin

Analysing

Preparation ~

Disciplines ~

About ~

Help ∨

ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D

Project content

In the third curation project of the <u>CLARIN-D working group 1 "German Philology" (F-AG 1)</u> an existing corpus of computer-mediated communication (CMC), the Dortmund Chat Corpus, and samples of other CMC resources will be restructured to conform to current standards for the representation of corpora in the Digital Humanities context. The main goal of this work is to pave the way for the inclusion of linguistically annotated CMC resources into CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology. To this end, the project will (1) transform the metadata and the annotations of the chat corpus into a TEI-compliant format, (2) enrich the data by further linguistic annotations, and (3) integrate the resulting resource into the CLARIN-D Corpus Infrastructures at the <u>Institute for the German Language (IDS)</u> and the <u>Berlin-Brandenburg Academy of Sciences (BBAW)</u>.

Duration: 05/2015 - 02/2015

Project team:

- Michael Beißwenger (U Dortmund)
- Angelika Storrer (U Mannheim)
- Eric Ehrhardt (U Mannheim)
- Axel Herold (CLARIN-D center at BBAW Berlin)
- Harald Lüngen (CLARIN-D center at IDS Mannheim)
- + other colleagues from IDS + BBAW providing support & expertise for specific work packages

Levels of customization

"Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned."

- customization₁ = change existing models so that they fit for
 your purpose ((adapt the standard to your
 needs))

The CLARIN-D Schema

Customization₂:

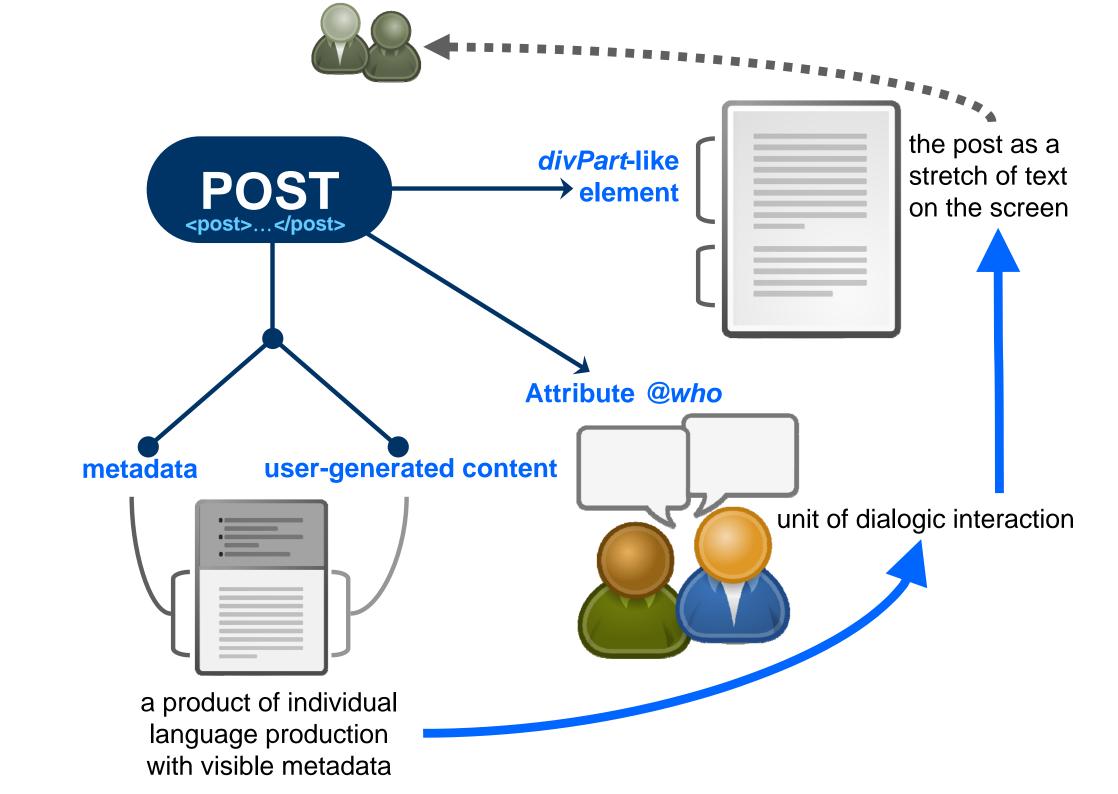
and extending the <post> model with additional attributes).

Customization₁:

Modification of standard models which are used more flexible in CMC than in traditional written genres (e.g. letters) (building on and extending suggestions from the DeRiK schema).

Customization₀:

Definition of best practices for the use of standard models (without modification) – e.g.: use of <w> for the representation of PoS information.



Extending the <post> model with attributes

post metadata which is automatically generated and displayed by the system:

@correspAction

is used to specify the 'sent|delivered|read' status a post as given for posts in whatsapp dialogues.

Legal values are: sent | delivered | read

cf. the element <correspAction> in the TEI standard which is used for a structured description of the place, the name of a person/organization and the date related to the sending/receiving of a message or any other action related to the correspondence [TEI module *correspondence description*:

Extending the <post> model with attributes

post metadata preserving information about what the author or other users have done with the original post after it has been initially posted (= its revision history):

@revisedBy

is used to refer to a person-ID in listPerson in the teiHeader and to specify the respective person as the one who created a revised version of the post by editing its original content.

@revisedWhen

is used to refer to a date in the timeline as the date when a revision has been performed.

Extending the <post> model with attributes

post metadata giving information that the researcher has reconstructed through interpretation and analysis:

@replyTo

indicates to which previous post the current post replies or refers to.

⇒ optional attribute which can be used for the annotation of thread structures

Two features of the CLARIN-D schema in more detail

- Content models for a flexible usage of structural elements of letters and CMC messages
- 2. Attributes to mark system-generated content

Structural elements of letters

Lettre XXII

De Madame de Tourvel à Madame de Volanges

J'ai l'honneur d'être, Madame, etc.

P.S.- Mme de Rosemonde & moi nous allons, dans l'instant, voir aussi l'honnête & malheureuse famille, & joindre nos secours tardifs à ceux de M. de Valmont. Nous le mènerons avec nous. Nous donnerons au moins à ces bonnes gens le plaisir de revoir leur bienfaiteur ; c'est, je crois, tout ce qu'il nous a laissé à faire.

De..., ce 20 août 17

Source: https://fr.wikisource.org/wiki/Les_Liaisons_dangereuses/Lettre_22

Structural elements of letters

Lettre XXII

De Madame de Tourvel à Madame de Volanges

bienfaiteur ; c'est, je crois, tout ce qu'il nous a laissé à faire.

De..., ce 20 août 17

•••••	
J'ai l'honneur d'être, Madame, etc.	
P.S Mme de Rosemonde & moi nous allons, dar malheureuse famille, & joindre nos secours tardif mènerons avec nous. Nous donnerons au moins à	s à ceux de M. de Valmont. Nous le

Source: https://fr.wikisource.org/wiki/Les_Liaisons_dangereuses/Lettre_22

Structural elements of letters in the TEI

Lettre XXII

De Madame de Tourvel à Madame de Vola	inges	
	<closer></closer>	<pre><postscript></postscript></pre>
l'ai l'honneur d'être, Madame, etc.		
P.S Mme de Rosemonde & moi nous allor nalheureuse famille, & joindre nos secours nènerons avec nous. Nous donnerons au m pienfaiteur ; c'est, je crois, tout ce qu'il nou	s tardifs à ceux de l noins à ces bonnes	M. de Valmont. Nous le
De, ce 20 août 17		

Source: https://fr.wikisource.org/wiki/Les_Liaisons_dangereuses/Lettre_22

Structural elements of letters in CMC messages

Source: de.rec.garten, Usenet 5.10.2014

Frequently in Wikipedia discussions, E-Mail or Usenet news:

- Message starts before opener, continues after closer, signed etc.
- opener, closer etc. embedded in paragraph, even sentence

Structural elements of letters or CMC messages in the CLARIN-D Schema

Definition of the model *model.floatP.cmc*

- collects TEI elements for structuring correspondence:
 <opener>, <closer>, <signed>, <poststript>, <trailer>
- is added to the content models for , <s>, <quote>

Two features of the CLARIN-D schema in more detail

- Content models for a flexible usage of structural elements of letters and CMC messages
- 2. Attributes to mark system-generated content

@auto: a new attribute in att.global (type: data.xTruthValue)

@who: TEI attribute promoted to att.global

auto="true" who="#user"

content marked by the present element was auto-generated by the system, but user deliberately activated its insertion by typing a macro or clicking a button,

auto="true" who="#system"

content marked by the present element was auto-generated by the system uncontrolled by the user e.g. as a consequence of some other action

> Morgenstund hat Gold im Mund [...]

auto-generated content: Usenet news: user starts message as a reply to a previous message

McMike geht in einen anderen Raum: Kreuzfahrt

auto-generated message: Chat: user changes chat room

auto-generated signature: Wikipedia talk page: user inserted his/her signature including a timestamp



CMC Metadata

IRD: Social Media and CMC Corpora for the eHumanities October $23\,/\,24\,2015,\, Rennes\,(FR)$



Types of metadata

Many different views on *metadata* are possible.

We want / do not want:

- ✓ resource description≈ intrinsic properties
- ✓ relation to the "real world" ≈ extrinsic properties
- ✓ description of data (item) annotation



Resource description

- owner
- hosting institution
- ▶ identifier(s)
- resource type
- encoding
- size
- creation date
- revision history
- documentation
- ▶ . . .

This applies to all resources.

There is a solid framework within TEI.



Relation to the "real world": computer mediated communication

TEI caters for communication (to some extent):

▶ /TEI/teiHeader/fileDesc/sourceDesc/listPerson (module namesdates)

```
ListPerson>
 <person xml:id="system">
   <persName>system
 </person>
 <person xml:id="A01">
   <persName type="nickname">Raebchen
       persName>
 </person>
 <person xml:id="A02">
   <persName type="nickname">McMike</persName>
 </person>
</listPerson>
```



Relation to the "real world": computer mediated communication

TEI caters for communication (to some extent):

- /TEI/teiHeader/fileDesc/sourceDesc/listPerson (module namesdates)
- limitations regarding properties of persons,
 e. g. (assumed) gender, participation in communities, . . .
- //timeline (module linking)
 to maintain temporal order in non-linear communications



Relation to the "real world": computer mediated communication

TEI does not (yet) cater for computer mediation

► /TEI/teiHeader/fileDesc/sourceDesc/recordingStmt (module spoken) <recordingStmt> <recording> plattformName=Unicum plattformURL=http://www.unicum.de/chat recStart="HH-MM" recEnd="HH-MM" recBy=" unknown" TNOM="758" TNOT="4101" view="ineli26" <date>1998</date> </recording> </recordingStmt>



Relation to the "real world": computer mediated communication

TEI does not (yet) cater for computer mediation

- ► /TEI/teiHeader/fileDesc/sourceDesc/recordingStmt
- "legacy" CMC specific metadata
- limitations regarding properties of the system / platform / environment (COMERE) (platformName, platformURL, . . .)
- possible solutions within TEI:
 - /TEI/.../recordingStmt/environment/*
 - ► /TEI/.../recordingStmt/fs/f (feature structures)
 - rely on external metadata descriptions
 - rely on (human readable) documentation



Data (item) annotation

- ► TEI guidelines
- specific ODD
- documentation such as annotation guidelines



Summary and Outlook

- metadata modeling in TEI using a mixture of explicit and implicit semantics needs amendments (CMC module vs. customization)
- CLARIN metadata model in CMDI (component metadata infrastructure) using explicit semantics



CUSTOMIZING A TEI SCHEMA

TEI customizations are defined in so-called ODD Files:

"One Document Does it all"

TEI Consortium (2015): TEI: Customization. http://www.tei-c.org/Guidelines/Customization/



CUSTOMIZING A TEI SCHEMA

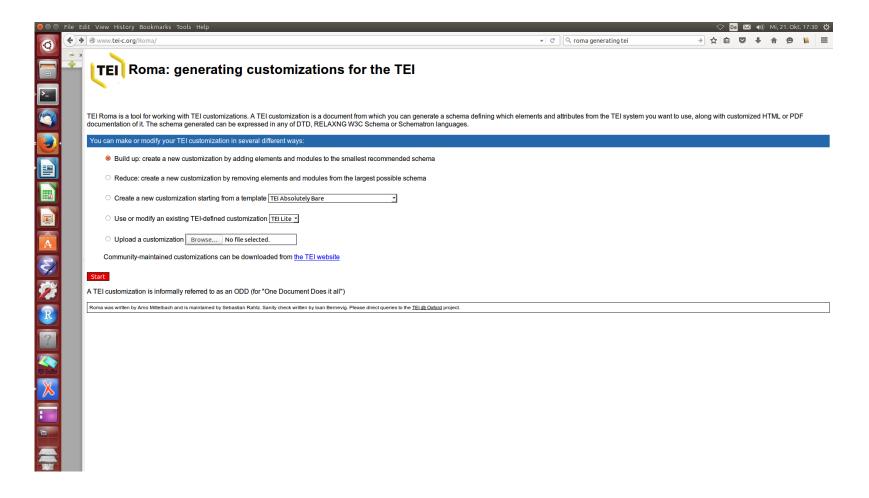
To prepare you own TEI customization, you can

- Start from scratch
- Start using an existing customization
 - a) as provided by the TEI, e.g. TEI light, TEI corpus
 - b) as provided by yourself or a third partye.g. the ODD provided by our project:

TEI CMC CHATCORPUS2CLARIN.odd

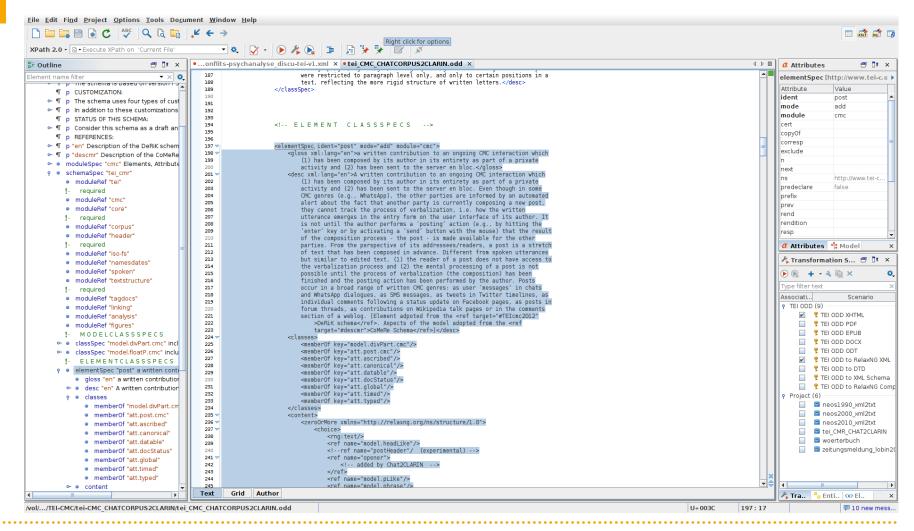


CUSTOMIZING A TEI SCHEMA HTTP://WWW.TEI-C.ORG/ROMA/





CUSTOMIZING A TEI SCHEMA OXYGEN XML EDITOR





USING A TEI SCHEMA FOR ANNOTATING CMC DOCUMENTS IN OXYGEN

To use a TEI customization for the (re-)annotation of CMC data, derive a schema, e.g. Relax NG, from the ODD file

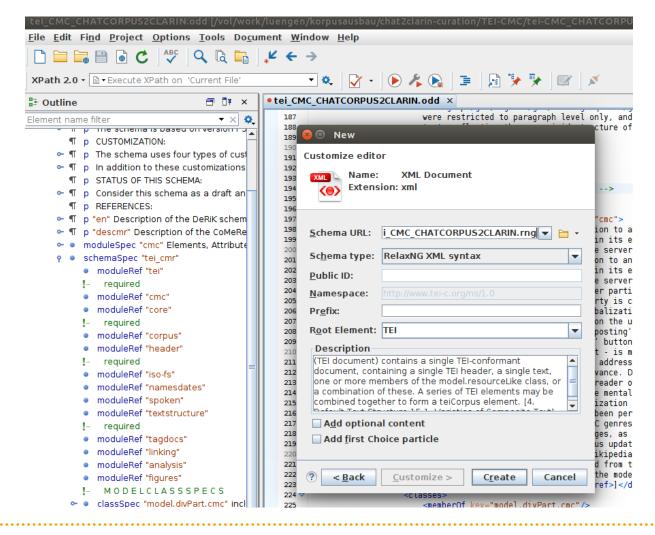
TEI_CMC_CHATCORPUS2CLARIN.rng

then

- Case 1: start a new TEI document from the schema
- Case 2: associate an existing TEI document with the schema



CASE 1: START A NEW TEI DOCUMENT FROM SCHEMA





USING A TEI SCHEMA FOR ANNOTATING CMC DOCUMENTS IN OXYGEN

To use a TEI customization for the annotation of CMC data, derive a schema, e.g. Relax NG, from the ODD file

TEI_CMC_CHATCORPUS2CLARIN.rng

then

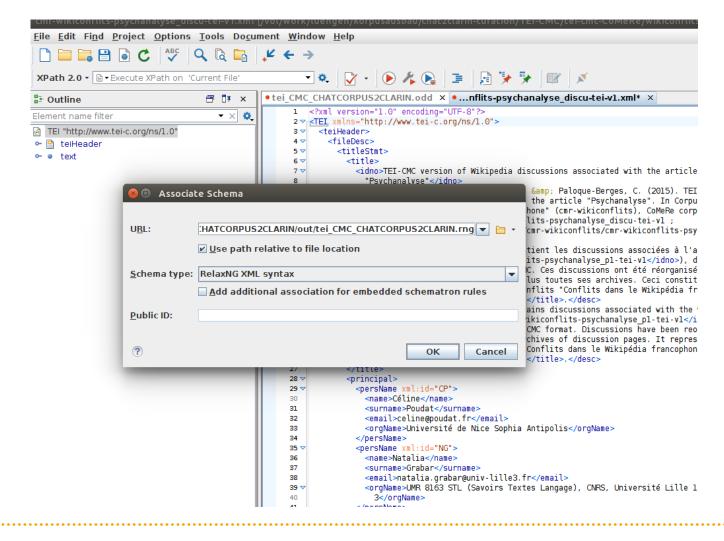
- Case 1: start a new TEI document from the schema
- Case 2: associate an existing TEI document with the schema



demo

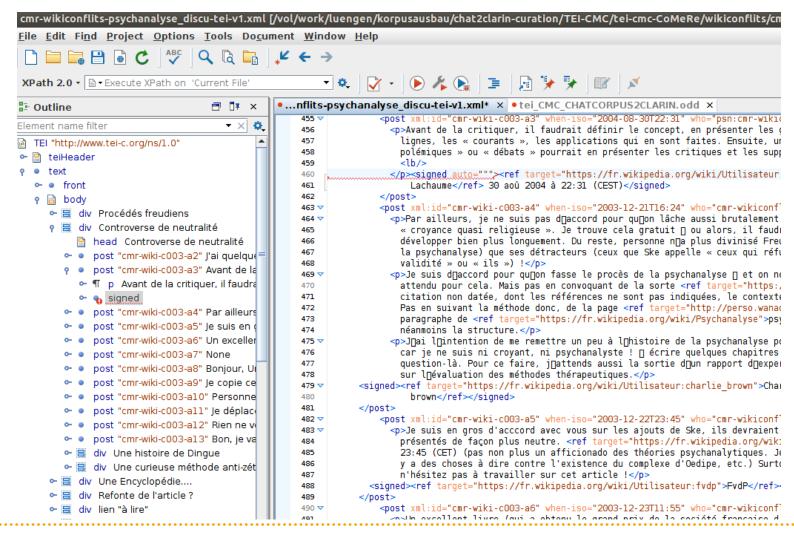


CASE 2: ASSOCIATE AN EXISTING TEI DOCUMENT WITH SCHEMA





CASE 2: ASSOCIATE AN EXISTING TEI DOCUMENT WITH SCHEMA





THANK YOU FOR YOUR ATTENTION

