Schemas and experiences from modeling German CMC Corpora in TEI

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

Project content

In the third curation project of the CLARIN-D working group 1 "German Philology" (FAG 1) 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 Institute for the German Language (IDS) and the Berlin-Brandenburg Academy of Sciences (BBAW).

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.”

customization$_0$ = use existing modeling concepts (henceforth: models) according to a specific practice which is useful for your purpose (adapt your needs to what the standard offers to you)

customization$_1$ = change existing models so that they fit for your purpose (adapt the standard to your needs)

customization$_2$ = add new models (extend the standard to fit for your needs)
The CLARIN-D Schema

Customization$_{2}$:

**Introduction of specific models for CMC-specific concepts:**<post> and <prod> for different types of user contributions to cmc interactions (building on the DeRiK and CoMeRe schema and extending the <post> model with additional attributes).

Customization$_{1}$:

**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).

Customization$_{0}$:

Definition of **best practices for the use of standard models** (without modification) – e.g.: use of <w> for the representation of PoS information.
the post as a stretch of text on the screen

POST

<post>...</post>

divPart-like element

Attribute @who

unit of dialogic interaction

metadata

user-generated content

a product of individual language production with visible metadata
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

1. Content models for a flexible usage of structural elements of letters and CMC messages

2. Attributes to mark system-generated content
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
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
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
Structural elements of letters in CMC messages

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>`, `<postscript>`, `<trailer>`

• is added to the content models for `<p>`, `<s>`, `<quote>`
Two features of the CLARIN-D schema in more detail

1. Content models for a flexible usage of structural elements of letters and CMC messages

2. Attributes to mark system-generated content
Attributes for marking system-generated content

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

@who: TEI attribute promoted to att.global

case 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
Am 03.04. 2015 um 05:46 schrieb Axel:
> Morgenstund hat Gold im Mund [...]
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

- Peter2 (Diskussion) 18:50, 22. Okt. 2008 (CEST)
CMC Metadata

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

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
Metadata

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.
Metadata

Relation to the „real world“:  
computer mediated communication  

TEI caters for communication (to some extent):  

▶ /TEI/teiHeader/fileDesc/sourceDesc/listPerson  
(module names dates)

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

▶ limitations regarding properties of persons,  
e.g. (assumed) gender, participation in communities, ...

▶ //timeline  
(module linking)  
to maintain temporal order in non-linear communications
Metadata

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
Metadata

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>
    <p>plattformName=Unicum
        plattformURL=http://www.unicum.de/chat
        recStart="HH-MM" recEnd="HH-MM" recBy="unknown"
        TNOM="758" TNOT="4101" view="ineli26"</p>
    <date>1998</date>
  </recording>
</recordingStmt>
Metadata

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
Metadata

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“

To prepare your 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 party
     e.g. the ODD provided by our project:

  [TEI_CMC_CHATCORPUS2CLARIN.odd]
CUSTOMIZING A TEI SCHEMA
HTTP://WWW.TEI-C.ORG/ROMA/

TEI Roma: generating customizations for the TEI

TEI Roma is a tool for working with TEI customizations. A TEI customization is a document from which you can generate a schema defining which elements and attributes from the TEI system you want to use, along with customized HTML or PDF documentation of it. The schema generated can be expressed in any of DTD, RELAXNG, W3C Schema or Schematron languages.

You can make or modify your TEI customization in several different ways:
- Build up: create a new customization by adding elements and modules to the smallest recommended schema
- Reduce: create a new customization by removing elements and modules from the largest possible schema
- Create a new customization starting from a template
- Use or modify an existing TEI-defined customization
- Upload a customization

Community-maintained customizations can be downloaded from the TEI website.

A TEI customization is informally referred to as an ODD (One Document Does it all)

Roma was written by Arno Miltzoch and is maintained by Sebastian Rauth. Sanity check written by Jean Berneving. Please direct queries to the TEI@Götting project.
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