A TEI Schema for the Annotation of CMC Genres

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, Angelika Storrer
A TEI schema for the annotation of CMC genres

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, Angelika Storrer (2012):
« A TEI Schema for the Representation of Computer-mediated Communication »,
Journal of the Text Encoding Initiative
Issue 3 (November 2012): TEI and Linguistics,
URL: http://jtei.revues.org/476 | DOI: 10.4000/jtei.476
Schema files and encoding examples:
http://www.empirikom.net/bin/view/Themen/CmcTEI
To date, there are no standards for representing and annotating CMC data for the purposes of linguistic analysis and building linguistic corpora.

**Option I: design an XML schema for CMC from scratch**

- schema perfectly fits with the needs of the individual project
- schema is idiosyncratic, resource (corpus) is not interoperable with other resources
To date, there are no standards for representing and annotating CMC data for the purposes of linguistic analysis and building linguistic corpora.

**Option II:** Try to design a schema which is compliant with an existing standard in the field of eHumanities

- interoperability of resources
- compliance with existing standard restricts the freedom to design everything in a way that perfectly fits for the particularities of CMC discourse
(1) It should provide elements for the annotation of units which are often regarded as “typical” for language use on the web and which are of special interest for everybody who wants to compare linguistic features of CMC discourse with the language documented in text corpora (such as the DWDS corpora).

(2) It should provide a model that is adapted to the structural particularities of CMC discourse and that takes into consideration that the interlocutors’ contributions to conversations in forums, chats, in wiki and weblog discussions, etc. can neither be adequately described as utterances in spoken conversations nor as paragraphs in written documents (even though they share features with both).
What should an annotation schema for DeRiK provide?

(3) It shall allow for an easy (but reversible) anonymization of CMC data for purposes in which the annotated data shall be made available as a resource for other researchers or for the public.

(4) It shall provide all information and metadata which are necessary for using and referencing random excerpts from the data as references in a general language dictionary as well as in the results of a corpus query.

(5) As far as possible, its core structure should be oriented on surface phenomena of the written discourse represented in the original CMC documents and thus be as independent as possible from any specific theory of CMC discourse in order to

- allow for an automatized transformation of these source documents into the target format (⇒ important for annotating large corpora);
- allow for a use as a basic model / exchange format in as many projects as possible.
(6) It should be open for extensions by other researchers in the field of empirical CMC research or by corpus designers who want to adapt the schema for their own project purposes and – thus – be able to work as a first step into a general discussion within the community about how a general format for the annotation of CMC corpora should look like and what it should comprise.

**Decision for option II:** Try to design a schema which is compliant with an existing standard in the field of eHumanities

- Existing standards don’t include elements or partial schemata that would fit for CMC.
- If an existing standard is flexible enough to allow for modifications/extensions as a part of the standard, this may be a good basis for getting started with a standard-conformant annotation schema for CMC.
We decided to design a schema on the basis of the annotation framework provided by the Text Encoding Initiative (TEI).

- TEI is a *de facto* standard in the field of eHumanities ⇒ widely used interchange format for a variety of genres and document types (1st version of the TEI guidelines: 1990) ⇒ interoperability of resources
- The DWDS corpora are already represented in TEI.

The TEI framework allows for a flexible adaptation and extension to new genres and document types which are not covered by the existing version of the standard already:

"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 is a central aspect of TEI usage and the Guidelines are designed with customization in mind."
Some basic features of our schema

1) Basic modelling decision: *textstructure* or *transcribed speech* as the basis?

2) Basic structural element: The *posting*

3) Elements for typical „netspeak“ phenomena

4) Modelling of users / anonymisation
To build a TEI pizza, take...

- **the pizza dough**
  - the core structure and tagsets

- **a base of your choice**
  - base „modules“ defined in the TEI framework, e.g.:
    - text structure
    - transcribed speech
    - dictionaries
    - manuscripts
    - performance texts
    - dictionaries (...)

- **the toppings of your choice**
  - additional tagsets for specific classes of features including – optionally – individual customizations/extensions of the tagsets used
To build a TEI pizza, take...

- **the pizza dough**
  = the core structure and tagsets

- **a base of your choice**
  = base „modules“ defined in the TEI framework, e.g.:
    - text structure
    - transcribed speech
    - dictionaries
    - manuscripts
    - performance texts
    - dictionaries (...)

- **the toppings of your choice**
  = additional tagsets for specific classes of features including – optionally – individual customizations/extensions of the tagsets used
CMC shares characteristics both with text and spoken conversation:

- Just as *spoken conversation* (and different from *text*), it is *dialogic interaction* in which each communicative move creates/changes the context for follow-up moves.

- Just as *text documents*, CMC discourse is organized in (and displayed on the screen as) *sequences of stretches of written text*.

CMC as text or as speech?

Options for modelling the users’ written contributions to CMC conversations with elements of the TEI standard modules (TEI P5):

- as **divisions** or **paragraphs** (TEI module „textstructure“)
- as **utterances** (TEI module „transcribed speech“)
The individual users’ written contributions to CMC dialogues are different both from spoken utterances and from paragraphs in monologic texts:

- Different from spoken utterances, the production of people’s contributions to CMC dialogues is text-like: they are first (1) composed by their authors in their entirety, then (2) sent to the server, then (3) shown on the screen as written messages before (4) they can be read and replied by other interlocutors.

- Under aspects of planning and coherence, the user contributions are similar to spoken utterances: there is no author who planned the entire dialogue in advance; instead, the dialogue is developed by the users as they go along with each posting creating the context for the partners’ next moves.

  ⇒ Even though the contributions are usually being displayed on the screen as individual paragraphs, the similarity with paragraphs of monologic texts is just a formal, not a functional one.
The element *posting*

*posting*: a content unit that is being sent to the server “en bloc”.

- Postings are recognizable by their formal structure and can be annotated automatically, even if they have different forms and structures in the different CMC genres.

- In order to integrate the new element into the TEI framework, we related it with the class of “division-like” elements. *division (div)* is a part of the TEI standard module *text structure*.

→ enables us to describe the inner structure of postings with standard TEI elements for text structure, text design etc.

→ **decision**: TEI module *text structure* as the basis for our schema ... but reinterpretation of some of its core elements (⇒ customization) – e.g.:
  - we added an attribute `@who` to be able to assign each posting to an author (which is not possible with divisions).
  - The attribute `@who` is borrowed from the TEI element *utterance (u)* from the TEI module *transcribed speech*.

⇒ The concept of the element *<posting>* represents the hybrid character of written user contributions to CMC dialogues (= a type of utterance between monologic text and turns in spoken conversation)
Annotation of selected „netspeak“ features

Above the posting level of CMC conversations, we describe types of **CMC macrostructures** (logfiles, threads as two main types of how postings can be structured in CMC documents).

Below the posting level, we describe selected linguistic features that appear in the content of postings (= **CMC microstructures**).

The current version of our schema includes a partial schema for the annotation of features which are often considered as being typical „netspeak“ elements:

- emoticons  :-D  ^^-^  O.O
- interaction words  *grins*  *freu*  /  *smirks*  *giggles*  *lol*
- addressing terms  @maria:  …  2angelika:  …  @all:  …
- interaction templates  😊  🌺  😄

Under a functional and syntactic perspective, these units behave pretty similar to interjections (aka **inserts**, **discourse particles**). In our schema, we build on the category „**Interaktive Einheiten**“ (**interaction signs**) introduced by the „Grammatik der deutschen Sprache“ (Zifonun et al. 1997) and describe these units as an extension of that category.
We subsume emoticons, interaction words and addressing terms under the category „interaction sign“ (which originally comprises interjections and responsives) because they share the following functional and syntactic characteristics:

**Function:**
They are used to express reactions to a partner’s utterance or to display the author’s emotions.

**Syntactic behaviour:**
They are not integrated in the sentence structure. *Instead:*
- they can occur in front of or after the sentence boundaries or as insertions („thrown between“ < lat. *interiectio*) *or*
- they are used „standalone“ as sentence-equivalent units in order to perform a speech act / a conversational move.

Typology of interaction signs

Our schema includes elements for all types of interaction signs as well as a range of attributes for the (semantic and functional) subclassification of these.
Anonymization issues

- Reasons
- Method
- Structure of the user profile
- Linking of the profiles and postings
Es mangelt nicht an einer Quelle, sondern es bestehen Zweifel, dass diese Darstellung so korrekt ist. –GiordanoBruno 18:25, 24. Nov. 2011 (CET)
A list of user profiles – first step of anonymization

<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
   <title>A list of user profiles – first step of anonymization</title>
</head>
<body>

<p>&lt;listPerson&gt;</p>
<p>&lt;person xml:id="A034" sex="2"> &lt;persName&gt;Thyes&lt;/persName&gt;</p>
<p>&lt;signatureContent&gt; &lt;ref target="http://de.wikipedia.org/wiki/Benutzerin:Thyes">Myriam Thyes&lt;/ref&gt;&lt;/signatureContent&gt;</p>
<p>&lt;/person&gt;</p>

<p>&lt;person xml:id="A035"> &lt;persName&gt;GiordanoBruno&lt;/persName&gt; &lt;signatureContent&gt;&lt;ref target="http://de.wikipedia.org/wiki/Benutzer:GiordanoBruno">GiordanoBruno&lt;/ref&gt;&lt;/signatureContent&gt;</p>
<p>&lt;/person&gt;</p>

<p>&lt;person xml:id="A036"> &lt;persName&gt;UW&lt;/persName&gt; &lt;signatureContent&gt;&lt;ref target="http://de.wikipedia.org/wiki/Benutzer:UW">Uwe&lt;/ref&gt;&lt;/signatureContent&gt;</p>
<p>&lt;/person&gt;</p>
&lt;/listPerson&gt;
</body>
</html>
Es mangelt nicht an einer Quelle, sondern es bestehen Zweifel, dass diese Darstellung so korrekt ist. --<autoSignature/>

Per Umweltschützen. Das hat in dem Artikel aber auch mal gar nichts zu suchen! --<autoSignature/>
Pilot corpus

- contains 90 discussions collected from Wikipedia talk pages
- has temporal restrictions
- has quantitative restraints
Conversion steps

1) Downloading the discussion
2) Converting HTML discussions into our schema:
   - Python Scripts
   - Manual processing

Result:

- separation of divisions, postings, heads of the divisions
- creation of user profiles
- distributing of some attributes
Open issues and future work

- **Current work:** Automatization of conversion of source data (Wikipedia discussions) into a basic version of our schema (annotation of postings, thread structure, user modelling).

- **Future work:** Creating cascading scripts which process an entire HTML source file with the source data (= one Wikipedia talk page as it was retrieved from the web) and produce a TEI document out of it.

- **Future work:** Adapt schema and conversion procedure to exceptions and problematic cases that occur during the automatization process.

- **Future work:** Proceed with other CMC genres (forum threads, weblog discussions, chat and instant messaging logfiles, Twitter timelines). Extend and enhance the schema during this process, where necessary.

- **Open issue:** To what extent can the annotation of features on the microlevel of postings (interaction signs) be automatized?
Open issues – beyond the DeRiK project context

- The schema developed for DeRiK is open for extensions and modifications by other researchers in the field of empirical CMC research or by corpus designers who want to adapt the schema for their own project purposes (≡ TEI conventions for customization).

- Beyond its use for the annotation of the DeRiK data, the schema is intended as a first step into a general discussion within the community about what a general format for the annotation of CMC data should (could) look like and what it should include.
The TEI “pizza model” – future perspective?

To build a TEI pizza, take...

- **the pizza dough**
  = the core structure and tagsets

- **a base of your choice**
  = base „modules“ defined in the TEI framework:
    - text structure
    - computer-mediated discourse
    - transcribed speech
    - dictionaries
    - manuscripts
    - performance texts
    - dictionaries (…)

- **the toppings of your choice**
  = additional tagsets for specific classes of features including – optionally – individual customizations/extensions of the tagsets used
<thread>
  <posting> ... </posting>
</thread>

Thank you for your attention! :-)

Michael Beißwenger: michael.beisswenger@tu-dortmund.de
Maria Ermakova: ermakovam@googlemail.com
Alexander Geyken: geyken@bbaw.de
Lothar Lemnitzer: lemnitzer@bbaw.de
Angelika Storrer: angelika.storrer@tu-dortmund.de