CMC-core: A basic schema for encoding corpora of computer-mediated communication in TEI
Michael Beißwenger, Harald Lüngen, Laura Herzberg, Ciara R. Wigham

To cite this version:
Michael Beißwenger, Harald Lüngen, Laura Herzberg, Ciara R. Wigham. CMC-core: A basic schema for encoding corpora of computer-mediated communication in TEI. 7th Conference of Computer-Mediated Communication and Social Media Corpora, Sep 2019, Cergy, France. halshs-02305756

HAL Id: halshs-02305756
https://halshs.archives-ouvertes.fr/halshs-02305756
Submitted on 4 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CMC-core: A basic schema for encoding corpora of computer-mediated communication in TEI

Michael Beißwenger, Universität Duisburg Essen, DE
Harald Lüngen, Leibniz-Institut für Deutsche Sprache, Mannheim, DE
Laura Herzberg, Universität Mannheim, DE
Ciara R. Wigham, Université Clermont Auvergne, FR

Motivation

A standard for the TEI representation of CMC corpora is needed for three purposes:
- as a prerequisite for the exchange, interconnection, and combined analysis of CMC corpora of different origins, different languages, and different genres — i.e. for the interoperability and sustainability of CMC corpora,
- to facilitate the merging and combined analysis of CMC corpora with corpora of other types, namely text corpora and spoken language corpora,
- for the integration and exploitation of CMC corpora in existing language resource infrastructures and with established tools.

Basic units of CMC discourse

<table>
<thead>
<tr>
<th>CMC unit</th>
<th>type of data</th>
<th>TEI-PS model</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoken / multimodal</td>
<td>transcription of speech</td>
<td>&lt;&lt;utterance&gt;&gt;</td>
</tr>
<tr>
<td>bodily activity</td>
<td>textual description</td>
<td>&lt;&lt;kinesic&gt;&gt;</td>
</tr>
<tr>
<td>onscreen activity</td>
<td>transcription or textual description</td>
<td>&lt;&lt;incident&gt;&gt;</td>
</tr>
<tr>
<td>post</td>
<td>(function: perform next move in a sequential interaction)</td>
<td></td>
</tr>
</tbody>
</table>

Temporal (and resulting pragmatic) preconditions of post CMC instantiation: chat

The TEI special interest group Computer-Mediated Communication

Goal (2013)
- provide encoding schemas for representing CMC corpora, compatible with the TEI framework (as ‘TEI customisations’)

Goal (2017)
- develop the schemas further and transform them into a TEI Feature Request, making an official proposal for an extension of the TEI standard with specific models for CMC

Results (until 2019)
- Three schemas (TEI customisations), used for encoding CMC corpora for a range of projects and genres
  - DerIK schema: Beißwenger et al. 2012
  - CoMeRe schema: Chanier et al. 2014
- Assessment of encoding experiences at TEI, DARIAH and CLARIN events, as part of the German DFG network Empirkom and the French CoMeRe network, at the CMC-CORPORA conferences and at a workshop on standards for CMC corpora.
- The CMC-core schema unifying models of all three previous schemas applying a “reduce to the max” maxim (2018/19) ⇒ Basis for a Feature Request to the TEI Council to implement these models as part of the official standard in late 2019.

CMC-core is encoded as an ODD that extends the official TEI by introducing four types of specifications:

1. The new module cmc: groups the new CMC-specific features so that they can be selected or deselected for a schema at once.
2. The new model class model.divPart.cmc: contains <<post>> and makes it available on the divPart level, i.e. allows for combining <<post>>, <<kinesic>, <<incident>> on the same level (Listing 4).
3. The new element <<post>> along with the three new, <<post>>-specific attributes @mode, @replyTo, and @indentLevel (Listings 1.2).
4. The new attribute class att.global.cmc containing the new, global attribute @creation for encoding how the text content was created in the CMC environment.

Possible values: “human”, “template”, “system”, “bot”, “unspecified”.

References


Encoding examples

Listing 1: Discussion thread on a Wikipedia talk page

Listing 2: Written and spoken post in WhatsApp chat interaction, from MoCoDa2 (2018)

Listing 3: Blog comment replying to a previous comment, from the Scilogs corpus (Gruntm Suárez et al. 2016).

Listing 4: Second Life multimodal chat example, adapted to CMC-core, from Chanier & Wigham (2015).