



HAL
open science

CMC-core: A basic schema for encoding corpora of computer-mediated communication in TEI

Michael Beisswenger, Harald Lungen, Laura Herzberg, Ciara R. Wigham

► To cite this version:

Michael Beisswenger, Harald Lungen, 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://shs.hal.science/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.

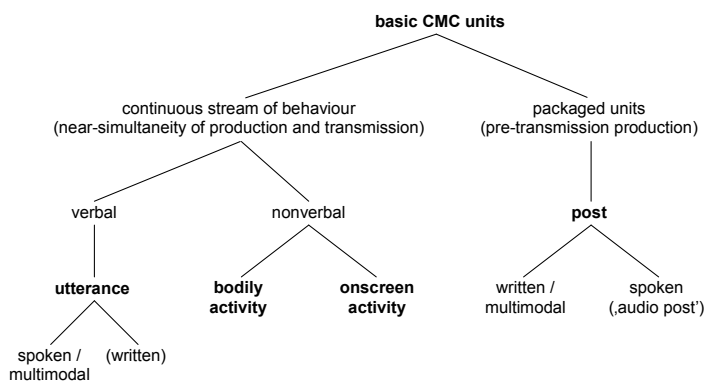
CMC-core: A basic schema for encoding corpora of computer-mediated communication in TEI

Motivation

A standard for the TEI representation of CMC corpora is needed for three purposes:

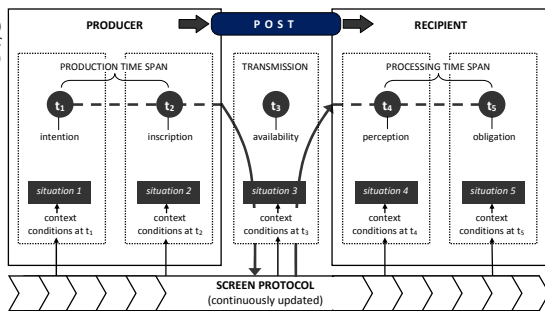
- o as a prerequisite for the exchange, interconnection, and combined analysis of CMC corpora of different origins, different languages, and different genres – i.e. for the interoperability and sustainability of CMC corpora,
- o to facilitate the merging and combined analysis of CMC corpora with corpora of other types, namely text corpora and spoken language corpora,
- o for the integration and exploitation of CMC corpora in existing language resource infrastructures and with established tools.

Basic units of CMC discourse



CMC unit	type of data	TEI-P5 model
utterance (spoken / multimodal)	transcription of speech	<u> (utterance)
bodily activity	textual description	<kinesic>
onscreen activity	transcription or textual description	<incident>
post	a stretch of text submitted to the server at once (function: perform next move in a sequential interaction)	?

Temporal (and resulting pragmatic) peculiarities of post-based CMC (instantiation: chat)



The TEI special interest group Computer-Mediated Communication

Goal (2013)

- o provide encoding schemas for representing CMC corpora, compatible with the TEI framework (as 'TEI customizations')

Goal (2017)

- o 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)

- o Three schemas (TEI customizations), used for encoding CMC corpora for a range of projects and genres
 - *DeRik schema*: Beißwenger et al. 2012
 - *CoMeRe schema*: Chanier et al. 2014
 - *CLARIN-D schema*: Lungen et al. 2016
- o Assessment of encoding experiences at TEI, DARIAH and CLARIN events, as part of the German DFG network Empirikom and the French CoMeRe network, at the CMC-CORPORA conferences and at a workshop on standards for CMC corpora.
- o 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 in a nutshell

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>, <u>, <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

- Beißwenger, Michael, 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 3. <http://jtei.revues.org/476> (DOI: 10.4000/jtei.476) (30.08.2019).
- Beißwenger, Michael, Lungen, Harald (2019, under review): *CMC-core*: a schema for the representation of CMC corpora in TEI. Submitted for: Céline Poudat, Ciara R. Wigham & Loïc Liégeois (eds.): *Corpus complexes. Traitements, standardisation et analyse des corpus de communication médiée par les réseaux*.
- Chanier, Thierry, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi & Djamel Seddah (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal of Language Technology and Computational Linguistics* 29 (2), 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf (30.08.2019).
- Lungen, Harald, Michael Beißwenger, Axel Herold & Angelika Storrer (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In Stefanie Dipper, Friedrich Neubarth & Heike Zinsmeister (Eds.), *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 1561–64.

Encoding examples

```
<div type="thread">
<head>Naturally occurring?</head>
<post mode="written" xml:id="p4" indentLevel="0" who="#u005"
synch="#u005">
<p>I'm not sure that this is a proper criterion, or even what this means.
What if we set an explosion that breaks a comet into two pieces? What if
we build a moon? Cheers, <signed creation="template"><ref
target="/wiki/User:Greenodd">Greenodd</ref> <ref
target="/wiki/User:talk.Greenodd">talk</ref> <time>01.00, 21
July 2011 (UTC)</time></signed>
</p>
</post>
<post mode="written" xml:id="p5" indentLevel="1" replyTo="#p4"
who="#u006" synch="#u006">
<p>Those haven't happened. If they do, we can revisit the concern. <signed
creation="template"><ref target="/wiki/User:Praemonitus">
Praemonitus</ref> <ref target="/wiki/User:talk.Praemonitus">
talk</ref> <time>01.15, 1 April 2015 (UTC)</time></signed>
</p>
</post>
</div>
```

Listing 1: Discussion thread on a Wikipedia talk page

```
<post mode="spoken" creation="human" synch="#u003" who="#A05"
xml:id="m7"> Sagt Anne auch gerade. JA! Kann ich zustimmen. </post>
<post mode="written" creation="human" synch="#u003" who="#A02"
xml:id="m8"> Da kostet ein Haarschnitt 50 € <figure type="emoji"
creation="template">
<desc type="meaning">face screaming in fear</desc>
<desc type="unicode">U+1F631</desc>
</figure>
</post>
```

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

```
<post xml:id="p5" type="comment" who="#u4" synch="#u005" replyTo="#p4">
<p>Wenn Sie diesen Gruppen also "mangelnde Bildung" attestieren wollen,
so verwenden Sie bereits einen bestimmten, kulturgebundenen Bildungshe
griff.</p>
<p>Ich hoffe doch, wir können beim Bildungsbegriff der Aufklärung
bleiben. Wer das nicht möchte, hat die Wissenschaft
verlassen.</p>
</post>
```

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

```
<text>
<body>
<u xml:id="cmr-archi21-siref-es-j3-1-a191" who="#tingrabu"
start="cmr-archi21-siref-es-j3-1-ts373"
end="cmr-archi21-siref-es-j3-1-ts430">ok hm for me this
presentation was hm <pause dur="PT1S"/> become too fast because
it's always the same in our architecture school eh we have not
time and hm <pause dur="PT1S"/> too quickly sorry [...]
</u>
<kinesic xml:id="cmr-archi21-siref-es-j3-1-a192" who="#fromeoz"
start="cmr-archi21-siref-es-j3-1-ts376"
end="cmr-archi21-siref-es-j3-1-ts377" type="body"
subtype="kinesics">
<desc>
<code>eat<popcorn></code>
</desc>
</kinesic> [...]
</post mode="written" creation="human"
xml:id="cmr-archi21-siref-es-j3-1-a195" who="#tfrez2"
start="cmr-archi21-siref-es-j3-1-ts380"
end="cmr-archi21-siref-es-j3-1-ts381" type="chat-message">
<p>it went too quickly?</p>
</post> [...]
```

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