Originally published in: Longhi, Julien; Marinica, Claudia (eds.): Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019) 9-10 September 2019. Cergy-Pontoise University, France. - Cergy-Pontoise, France: Cergy-Pontoise University, France. Pp. 74-75

cmc-core: A basic schema for encoding CMC corpora in TEI

Michael Beißwenger[1], Laura Herzberg[2], Harald Lüngen[3], Ciara R. Wigham[4]
[1] University of Duisburg-Essen, Germany [2] University of Mannheim, Germany [3] Leibniz-Institute for the German Language, Mannheim, Germany [4] University Clermont-Auvergne, France
E-mail: michael.beisswenger@uni-due.de, herzberg@uni-mannheim.de, luengen@ids-mannheim.de, ciara.wigham@uca.fr

Abstract

Since 2013 representatives of several French and German CMC corpus projects have developed three customizations of the TEI-P5 standard for text encoding in order to adapt the encoding schema and models provided by the TEI to the structural peculiarities of CMC discourse. Based on the three schema versions, a 4th version has been created which takes into account the experiences from encoding our corpora and which is specifically designed for the submission of a feature request to the TEI council. On our poster we would present the structure of this schema and its relations (commonalities and differences) to the previous schemas.

Keywords: CMC, cmc corpora, standard, TEI

Poster abstract

In close interconnection with the activities of the CMC corpora community, since 2013 representatives of several CMC corpus projects have been developing customizations of the TEI P5 standard for text encoding in order to adapt the encoding schema and models provided by the TEI to the structural peculiarities of CMC discourse. Since the TEI-P5 standard does not offer any specific models for the representation of CMC discourse the goal of the group - which could install a special interest group (SIG) on CMC as part of the TEI community - was twofold:

(1) *short-term goal:* provide encoding schemas which people could use for representing CMC corpora in a way which is compatible with the general structure of TEI documents ('TEI customizations') even though the TEI standards does not include models of CMC.
(2) *long-term goal:* gather and evaluate experience from different corpus projects using these schemas; develop the schema further and transform it into a 'feature request' to make an official proposal for an extension of the TEI standard with specific models for CMC.

The SIG started from a 1st schema draft (Beißwenger et al. 2012, 'DeRiK schema') which formed the basis for the creation of an extended schema by the French CoMeRe group (Chanier et al. 2014, 'CoMeRe schema') which was used for the encoding of 14 French CMC corpora. The latter was further developed in the German ChatCorpus2CLARIN project 2015/16 and adopted for encoding German chat, Wikipedia and Usenet corpora (Lüngen et al. 2016, Beißwenger 2018, 'CLARIN-D schema').
Based on the three schema versions, a 4th version has been created in 2018 which takes into account the experiences from encoding the abovementioned French and German CMC corpora and which is specifically designed for the submission of a feature request to the TEI council. On our poster we would present the structure of this schema and its relations (commonalities and differences) to the previous schemas.
The goal of the new schema, dubbed *cmc-core*, is to reduce the previous schema drafts "to the max" and provide an essential architecture of concepts which are needed for the representation of documents which typically form the basis of every CMC corpus. *Cmc-core* provides <post> as a basic model to describe the peculiarities of user contributions to CMC interactions which are - even in the case of spoken "audio posts" in WhatsApp sequences - characterized by a temporal rupture between production and transmission which makes them different from turns in spoken interactions. Instances of posts are constituents of 'CMC macrostructures' (logfiles or threads) which are represented using the element from the TEI standard. Posts can be subclassified by several attributes:

- For the distinction of spoken vs. written posts, we introduced the @mode attribute with its two possible values "written" and "spoken".
- For encoding a technical back reference from one post to one previous post, and the indentation level of wiki talk contributions, we use the attributes @replyTo and @indentLevel which were already included in the previous schema drafts.
- For classifying content according to different types of creators ("human", "template", "system", "bot", "unspecified") we use the attribute @creation (a further development of the attribute @auto from the previous schema).

The new cmc-core schema will be made available together with sample encodings of chat, twitter, wiki talk and transcribed 2nd life interactions on the CMC-SIG pages in the TEI wiki in August 2019 (https://wiki.tei-c.org/index.php?title=SIG:CMC). After publication in the TEI wiki members of the TEI-CMC-SIG and colleagues from the CMC corpora community will be invited (via their mailing lists) for critical review and comments. It is planned to submit the feature request to the TEI community by October 2019.

```
<post key="1043796093786566656"
      mode="written"
      creation="human"
      type="tweet"
      subtype="retweet"
      who="#u2"
      synch="#tweetsbcrn18.t003"
      xml:lang="de"
      xml:id="p5">
    <time creation="system"> 14:35 </time> Immer wieder gerne. Kann ich mich schon für nächstes
    Jahr als Empfangs- <ref type="hashtag" target="https://twitter.com/hashtag/Engel?src=hash"
        >#Engel</ref> für das nächste BarCamp bewerben <figure type="emoji" generation="templat
            ><desc type="meaning">zany face</desc>
        <desc type="unicode">U+1F92A</desc></figure>
    <ref type="hashtag" target="https://twitter.com/hashtag/bcrn18?src=hash">#bcrn18</ref>
    <trailer>
        <!-- The following is CoMeRe Style  -->
        <fs>
            <f name="favoritecount">
                <numeric value="4" />
            </f>
        </fs>
    </trailer>
</post>
```

Figure 1: Encoding of a tweet according to *cmc-core*

References

Beißwenger, M. (2018): Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI. In: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hg.): Digitale Infrastrukturen für die germanistische Forschung. (= Germanistische Sprachwissenschaft um 2020 6). Berlin u.a., pp. 307-349.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. (2012): A TEI Schema for the Representation of Computer-mediated Communication, *Journal of the Text Encoding Initiative*, Issue 3. DOI : 10.4000/jtei.476

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J. & Seddah, D. (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics. *JLCL (Journal of Language Technology and Computational Linguistics),* Issue 2, pp. 1-31

Lüngen, H., Beißwenger, M., Ehrhardt, E., Herold, A. & Storrer, A. (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016).* Ruhr-Universität Bochum.