Originally published in: Corpus [Online] (2020) no. 20, 24 pp.
DOI: *https://doi.org/10.4000/corpus.4549*

CMC-core: a schema for the representation of CMC corpora in TEI

Le CMC-core *: un schéma de représentation des corpus de la CMR en TEI*

Michael Beißwenger and Harald Lüngen

1. Introduction

1 In this paper we describe a schema and models which have been developed for the representation of corpora of computer-mediated communication (CMC corpora) using the representation framework provided by the Text Encoding Initiative (TEI). The schema presented here is the result of the activities and discussions within an international community of researchers who have been building, annotating and processing CMC data for the integration into corpus infrastructures (CLARIN, ORTOLANG) and use these corpora for purposes of linguistic research on linguistic variation and language change in and through the impact of internet-based communication technologies and applications. Discourse in the scope of CMC corpora (= "computer-mediated communication") is characterised as *dialogic, sequentially organised interchange between humans* which is conducted using

- communication technologies such as chats, messengers, online forums;
- social media platforms and applications such as Twitter, Facebook, Instagram or WhatsApp;
- the communication functions of collaborative platforms and projects (e.g. in the Wikipedia or in learning environments);
- or 3D environments (e.g. Second Life, gaming environments).

2 Discourse found in CMC exhibits features that cannot be adequately handled by schemas and tools developed for the representation, annotation and processing of discourse that conforms to the written standard and the structural conventions of established text types (e.g., newspaper articles, prose, scientific articles). It also significantly differs from the language and structure of spoken conversation so that

models for the representation of spoken language cannot be adopted for the representation of CMC corpora without critical revision, either.

3 Nonetheless, a solution for the representation of CMC discourse which complies with established standards for the representation of language corpora is a great desideratum. Such a solution is needed for three purposes:

i. 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;

ii. to facilitate the merging and combined analysis of CMC corpora with corpora of other types, namely text corpora and spoken language corpora;

iii. for the integration and exploitation of CMC corpora in existing language resource infrastructures and with established tools (CLARIN, Beißwenger et al. 2017).

4 Since 2013, the special interest group (SIG) "computer-mediated communication" in the Text Encoding Initiative (TEI) has worked towards such a solution: It has created four customised TEI schemas which build on existing models for text genres from the official TEI encoding framework TEI P5 and adapt these models for the representation of CMC data of different origin covering a range of prominent CMC genres. The four schemas have not been developed independently; instead, the schemas mark milestones on a path on which the previous schema, and the lessons learnt in using it for the representation of corpora, provided the basis for the development of the next schema in line. The schema developed in the context of a planned German reference corpus on CMC published in the TEI Journal ('DeRiK schema', Beißwenger et al. 2012) as well as its variant adapted for representing a German Wikipedia corpus (Margaretha & Lüngen 2014) marked the initial points in that process. The French CoMeRe group around Thierry Chanier adapted and extended the DeRiK schema to represent 14 existing French CMC corpora on different CMC genres in a uniform and interoperable way ('CoMeRe schema', Chanier et al. 2014). Whereas the two variants of the DeRiK schema were focused on writing-based CMC (chats, forums, tweets, Wikipedia), the CoMeRe schema extended its scope by introducing additional models for the representation of forms of multimodal CMC as represented in the French LEarning and TEaching Corpora (LETEC, Reffay et al. 2009) or to be observed in platforms such as Second Life (Wigham & Chanier 2013). Experiences with using the DeRiK and CoMeRe schemas for modelling CMC discourse in corpora were intensively discussed in activities and meetings of the CMC-SIG held as part of DARIAH and TEI events in Rome (2013, 2014) and Lyon (2015) as well as in the context of conferences of the German DFG-funded scientific network "Empirical analysis of internet-based communication" (Empirikom, http://www.empirikom.net) and the French Network "Communication mediée par les résaux" (CoMeRe, https://corpuscomere.wordpress.com/), and as part of the international conference series on CMC corpora (www.cmc-corpora.org) which originated from the cooperation of the German and French scientific networks. In 2016, a further developed version of the preceding schemas was made available as a result of the CLARIN-D curation project "ChatCorpus2CLARIN" ('CLARIN-D schema', Lüngen et al. 2016). This schema version was further discussed at an international workshop on standards for CMC and social media corpora held in Essen and, applying a "reduce to the max" maxim, subsequently transformed into a schema version which introduces a basic architecture and a set of TEI models needed for the representation of the most essential structural elements of CMC discourse. This schema –termed *CMC-core*– shall

provide the basis for a feature request (in preparation) which is planned to be submitted to the TEI Council and community in late 2019 as a suggestion for the implementation of models for the representation of CMC in a future version of the official TEI standard.

5 All four schema versions are available in the form of RNG schemas and ODD documents on the SIG pages of the TEI wiki (https://wiki.tei-c.org/index.php?title=SIG:CMC) for annotation work in other projects. The schemas have been tested and proven useful for the following CMC corpora and genres:

(1) *CoMeRe*[1] corpora: Under the supervision of Thierry Chanier, a set of 14 French CMC corpora was curated to be encoded in the customisation *CoMeRe-TEI* and subsequently published in the national ORTOLANG repository. The 14 corpora comprise the genres SMS, Twitter, email, text chat, wiki talk, weblog, discussion forum, and multimodal CMC (e-learning and *Second life*) and are available from https://repository.ortolang.fr/api/content/comere/v2/comere.htm (Wigham & Ledegen 2017, Poudat & Wigham in this volume).

(2) The *Dortmund Chat Corpus*, containing German language chat logs in various subgenres collected between 1998-2006, has been curated as part of a CLARIN-D project to be encoded in the CLARIN-D TEI customisation for CMC, to be POS-tagged and published in the CLARIN-D infrastructure (IDS and BBAW repositories) as *Dortmund Chat Corpus 2.2* (http://repos.ids-mannheim.de/fedora/objects/clarin-ids:chat2.2.000000/datastreams/CMDI/content, Lüngen et al. 2016).

(3) Wikipedia and Usenet corpora in DEREKO:

(3.1) *Wikipedia corpora in DEREKO*: Since 2013, Wikipedia corpora including Wikipedia talk pages have been built from Wikipedia dumps and included in the German Reference Corpus DEREKO. The TEI customisation I5, which is used for DeReKo, includes the DeRiK schema (Margaretha & Lüngen 2014). The same schema was also used to build Wikipedia corpora for other European languages, all available from the IDS repository (http://hdl.handle.net/10932/00-03B6-5583-B5A0-1201-0, Lüngen & Kupietz 2017).

(3.2) *Usenet news Corpora in DEREKO*: In a similar way, corpora from German language Usenet newsgroups have been downloaded from a news server, encoded in I5/the DeRiK schema and included in the German Reference Corpus DeReKo (http://www1.ids-mannheim.de/kl/projekte/korpora.html, Schröck & Lüngen 2015).

(4) *SciLogs corpus*: The CLARIN-D schema has also been used as markup for encoding the experimental German SciLogs blog corpus at the University of Gießen (Grumt Suárez et al. 2016).

(5) *MoCoDa2*: The corpus of German private chat communication from the ongoing project MoCoDa2 (Mobile Communication Database) will soon be converted in *CMC-core* (https://db.mocoda2.de/, Beißwenger et al. 2019).

6 In the following sections we will first give a rationale for why we support the idea to create a solution for the representation of CMC corpora using the TEI even though the current version of the TEI encoding framework and guidelines (TEI P5) does not offer any specific models for CMC. We will provide an overview of requirements a basic schema for CMC should fulfil (Sect. 2) and introduce the basic structural units that constitute written, spoken, and multimodal CMC discourse from a linguistic point of view (Sect. 3). In Sect. 4 we will discuss to what extent models from TEI P5 can be adopted for the modelling of CMC data and which extensions are required to "CMCify the TEI" so that it can provide a schema which is practical for the representation of CMC data in corpora. In Sect. 5 we will describe the extensions to TEI P5 provided by the *CMC-core* schema from a text-technological point of view and illustrate these model extensions with the help of encoding examples from different CMC corpora. The article

will close with an outlook on the further path towards the intended "CMCification of the TEI".

2. Representing CMC in TEI: fundamental decisions and considerations

7 The annotation of a linguistic corpus is typically based on a model according to the type of discourse or linguistic phenomenon the corpus data is meant to represent. In order to establish the representation relation, segments of corpus data are labelled as instances of structural components which in turn are defined and characterised as parts of the model. The model itself is typically described using a formal modelling language (e.g. RNG), and the labels added to the corpus data are expressed following the conventions of a markup language (e.g. XML). One main objective of corpus annotation is to guarantee the sustainability and interoperability of the resource (*curation-driven modelling*, cf. Jannidis 2017: 102). *Sustainability* implies that the resource can be processed with standard tools and software-independently even decades after its creation; for that purpose it is crucial to represent the resource in a non-proprietary encoding format using well-documented encoding schemas. *Interoperability* means that the resource is structured in a way that it may be merged and combined with resources of other creators; for this purpose it is important to encode and represent the resource compliant with representation *standards* which have been established as acknowledged solutions for the modelling of resources in the respective scientific domain. Standards serve to prevent that every creator of a corpus has to "reinvent the wheel" (Lobin 2010: 107); they provide solutions that have proven useful and practical in a broad range of projects which, as a result, can exchange their data and tools.

8 The representation of corpus data using a markup language is not primarily a "technical" task. Even though the XML code that results from annotating data in TEI may appear "technical" for people who are not familiar with reading annotations of that type, the main challenge of transforming raw linguistic data into a representation format which includes a description of their structure is a *modelling* challenge. Following Jannidis (2017: 107–108), modelling can be regarded as the process where a subject's understanding from the humanities' perspective meets the competence to express this understanding using a formal set of modelling procedures. In the best case, this process may allow for new research questions and foster new research.

9 A great example of the development of data models that integrate the humanities' perspective with a technological concept of data modelling are the encoding guidelines curated by the *Text Encoding Initiative* (*TEI*, http://tei-c.org), which can be considered "one of the big success stories of Digital Humanities" (Jannidis 2017: 107). Since 1996, the TEI develops and continuously refines models for the annotation of textual data in the humanities in a community-driven way, with a Technical Council ensuring consistency and practicability. The representation of these models is based on the Extensible Markup Language (XML) as a non-proprietary encoding format for descriptive text structuring which can be read and processed by a broad range of corpus tools so that it can justifiably be regarded as an "ASCII for the 21st century" (Sperberg-McQueen 2018: 292).

10 A basic model for the structural representation of corpora of computer-mediated communication should support the ideals of sustainability and interoperability in the best possible way and exhibit the following features (cf. Beißwenger 2018):

1. It should build on established representation standards for language resources in the field of corpus linguistics and digital humanities.
2. It should describe its subject in a most generic way so that it can serve as a model for a basic structural annotation of CMC resources in many corpus projects (on different CMC genres, for different languages).
3. At the same time, it should be specific enough to cover the essential structural peculiarities of CMC data, especially when compared with genres of edited text and spoken conversation.
4. It should be able to generate annotations for the basic structural components of CMC documents automatically from the encoding and layout information given in the raw data (e.g. from HTML code).
5. It should support the generation of an anonymised view of the data.
6. It should enable queries using established corpus query languages and query tools.

11 So far, there is no established standard for the representation of CMC data in linguistic corpora. The current version of the TEI encoding framework (TEI P5) does not include models which could be used "off the shelf" for the annotation of chat and messaging logfiles, forum threads, tweets and Instagram posts with their follow-up messages, "talk pages" in the Wikipedia, or multimodal, audio-, video- or 3D-based exchanges within videoconferencing tools, learning and gaming environments or virtual worlds. On the other hand, the TEI is a huge hub of expertise and concepts for the modelling of diverse sorts of text genres and of transcriptions of speech, or, more generally speaking, the modelling of discourse. TEI allows users to adapt and extend the "official" version of the standard for their purposes and defines rules for *customising* TEI for use in domains which are not yet covered by the framework:

> 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. (https://tei-c.org/guidelines/customization/).

12 The domain of CMC is a recent and prominent example of an "area that the TEI has not yet envisioned". Several approaches to customise TEI P5 for the annotation of CMC corpora published in 2012, 2014 and 2016 (Beißwenger et al. 2012, Chanier et al. 2014, Margaretha & Lüngen 2014, Lüngen et al. 2016) and discussed at TEI conferences and members' meetings have proven that TEI P5 provides a useful platform for the definition of representation schemas for CMC. Nevertheless, as long as a model for the representation of CMC is available "only" in form of TEI customisations, the official TEI standard is not up-to-date to cover this very prominent domain of discourse which, in the past few years, has become a subject of studies in a broad range of disciplines (within and beyond the humanities).

13 It is high time to include models for the representation of CMC in the TEI encoding framework and thus "CMCify the TEI" for usage in an emerging field of interdisciplinary research and for the annotation of a new type of linguistic corpora. *CMC-core*, which we describe in the following sections, is a customised TEI schema

which we regard as a suggestion for consideration in a future version of the TEI encoding framework and guidelines.

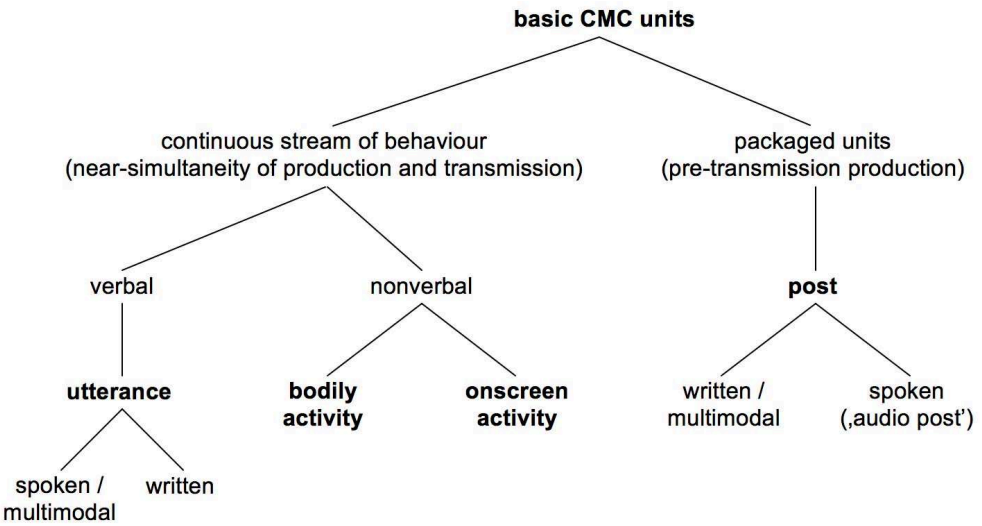
3. Basic units of CMC discourse

14 CMC is, by default, multimodal. Whereas early CMC systems (e.g. the prototypes of today's chat and messaging applications or Usenet developed in the 70s and 80s when bandwidth was still an essential factor for keeping transmission costs and capacities as low as possible) were completely ASCII-based, the lion's share of contemporary CMC technologies and applications allows for the combination of different modes (e.g. written or spoken language with icons, images, gestures) and even for the combined use of different CMC modalities in one and the same platform (e.g. combined use of an audio connection, a chat system and a 3D interface in which the users control a virtual avatar as in many MMORPGs or in virtual worlds such as Second Life).[2]

15 We refer to units of discourse which are produced by an interlocutor to contribute to an ongoing CMC interaction or joint CMC activity as *basic units of CMC discourse*. Contributions to an ongoing interaction are produced to perform a move as part of the further development of the interaction sequence, for instance in chats or forum discussions. Contributions to joint CMC activities may not directly be interactional but part of a collaborative project of the involved individuals, for instance editing activities in a shared text editor or whiteboard in parallel with an ongoing CMC interaction (chat, audio connection etc.) in the same CMC environment.

16 Basic units to CMC discourse can be described according to three criteria: (i) the temporal properties of how these contributions are produced by their creators, transmitted via CMC systems and made accessible for the recipients; (ii) whether the unit as a whole is realised in a verbal or nonverbal mode; (iii) for verbal units: whether the unit is realised in the written or spoken mode. A taxonomy of basic CMC units resulting from these criteria is given in Fig. 1.

Fig. 1. Types of contributions to CMC interactions



17 Some further elaborations on the taxonomy given in Fig. 1:

18 ○ The most important distinction refers to the temporal nature of units exchanged via CMC systems. The left part of the taxonomy describes units that are performed (= producer) and perceived (= recipient) as a continuous stream of behaviour and which are in most cases volatile. This means that they are not preserved as persistent units in a stored protocol the addressees could use to perceive and process units with some delay.[3] As a result, units which belong to the type 'continuous stream of behaviour' typically have to be perceived and processed simultaneously with their transmission, and in most cases transmission takes place simultaneously or at least near-simultaneously with their production.[4] Units of that type can be performed as

- *spoken utterances*, i.e. stretches of speech which are produced to perform a speaker turn in a conversation),
- *bodily activity*, i.e. nonverbal behavior (gesture, gaze) produced to perform a speaker turn or a backchannel from the recipient's position, either performed by the real body of an interlocutor (e.g. in a video conference) or performed through the virtual avatar of an interlocutor in a 3D environment,
- *onscreen activities*, i.e. non-bodily forms of behaviour that are transmitted to the group of interacting or coworking participants, for instance the editing of content in a shared text editor, etherpad or a shared whiteboard, which can be perceived by everybody simultaneously (as is the case in multimodal learning environments),
- *written utterances*, i.e. written language produced to perform a speaker turn which, different from units of the type *post*, is produced and transmitted to the addressees simultaneously (in a keystroke-by-keystroke mode). CMC systems in which written utterances are used as the standard unit for contributions (the most prominent one was *UNIX Talk*) can be considered historic as they do not play a relevant role in contemporary CMC anymore.[5]

19 ○ The right part of the taxonomy describes units in which the production, transmission and perception of contributions to CMC interactions are organised in a strictly consecutive order: The –verbal, nonverbal or multimodal– content of the contribution has to be produced before it is submitted to the system and is then presented on the screen as a preserved and persistent unit. We term this type of unit a *post*. Posts occur in two different variants:

- as *written / multimodal posts* which are produced with an editor form that is designed for the composition of stretches of written text; most of contemporary post-based CMC systems also offer users to include graphics and audiovisual content (emoji graphics, images, videos) into their posts and even to produce posts without verbal content; written / multimodal posts are the standard format for user contributions in primarily text-based CMC genres and applications such as chat, SMS, WhatsApp, Instagram, Facebook, Twitter, online forums or Wikipedia talk pages;[6]
- as "*audio posts*" that are produced using a recording function; in contrast to CMC units of the type *utterance* which are produced and transmitted simultaneously, "audio posts" first have to be recorded as a whole and are then submitted –as an audio file– to the CMC system; the availability of the recording is indicated in the screen protocol by a template-generated, visual post; the recipients can play the recording (repeatedly) by activating the play button displayed in the post on the screen. The most prominent CMC system that implements audio posts is the messaging application WhatsApp.

20 ○ Bodily activity is treated as a basic interactional unit of CMC discourse only in cases in which it is used to perform a move within the interactional sequence. It may also occur as a part of spoken, multimodal utterances (e.g. in video conferences); bodily

activity is then regarded as a nonverbal (and in these instances, co-verbal) part of the spoken, multimodal utterance.

21 Basic units of the types given in Fig. 1 constitute interactional and joint cooperative activity which takes place between two or more individuals in CMC environments. The model we will describe in Sect. 5 focuses on interactional activity, i.e. communicative activity with sequential organisation to which different participants contribute with basic units of the types *utterance*, *bodily activity* or *post*. The schema does not claim to provide a full and fine-grained set of models for a detailed description of joint cooperative activity that can be observed for instance in complex learning environments. The focus of the schema is on computer-mediated *communication*. Nevertheless, since CMC interactions may accompany joint cooperative activity, the schema includes a model for the representation of textual descriptions of *onscreen activity* in order to enable use even for the annotation of data from collaborative learning activities, as given e.g. in the French LEarning and TEaching Corpus *LETEC*, Reffay et al. 2009, Chanier et al. 2014).

22 Interaction in CMC environments is structured on two levels (cf. Beißwenger et al. 2012):

i. A *macrolevel* (or the *CMC macrostructure*) which is constituted by instances of the basic units *utterance*, *bodily activity* and *post* introduced above. Depending on the type(s) of units used in a CMC interaction or joint CMC activity, different types of macrostructures occur: In utterance-based CMC systems (i.e. in audio or video conferences) the macrostructure is negotiated by the speakers and organised on a turn-taking basis (with some limitations resulting from the fact that audio or video conferences mediated via the WWW are not always 100% as simultaneous as natural, non-mediated conversations). In post-based CMC systems, the macrostructure, i.e. the order and presentation of posts in the document on the screen, are the result of user activities *and* system routines; common macrostructure types in post-based systems are the *logfile,* in which user contributions are given in a chronological order, and the *thread,* in which chronological and topical structuring are combined.

ii. Different from CMC macrostructures, which are negotiated by the interlocutors or the result of user activities and system routines, we use the term *CMC microstructures* (or *microlevel*) for the description of the user-generated content of the CMC units, i.e. realisations of spoken or written language, optionally comprising multimodal elements as well, produced by one single interlocutor when performing an utterance or a post, or the forms of behavior produced by an interlocutor when performing a nonverbal contribution to a CMC interaction. The structural elements on the microlevel are therefore those elements that can be found in the content of individual users' contributions to CMC interactions while the constituting structural elements of the macrolevel are the users' contributions themselves. Structures on the microlevel are made of linguistic units and/or nonverbal units, hyperlinks, images and videos. The microstructure of a contribution to a CMC interaction is typically the result of the planning and production activity of one single author.

4. Why we need extensions to TEI P5 for the modelling of CMC

23 For the representation in corpora, CMC interactions have to be recorded and made available as documents with textual content. In the case of post-based CMC, the logfiles created by the server of the CMC system and displayed (as HTML) on the interlocutors' screens, together with the graphics, video and audio files that have been embedded,

can be used as raw data. In the case of utterance-based CMC, the audio or video conference has to be recorded and transcribed; the resulting textual descriptions represent the raw corpus data which then become subject to further structural annotation. In the case of activity from complex CMC environments in which the participants or collaborators use post-based and utterance-based CMC modes (e.g. an audio connection together with a chat), and probably a shared text editor or whiteboard in parallel, the data from the different modes have to be treated differently (chat → logfile, spoken conversation and onscreen activities → transcription) and then be represented and aligned to each other in the corpus document.

4.1 The macrolevel

24 Three of the four basic CMC units described in Sect. 3 can be represented with models from standard TEI. Even though these models have not been developed with CMC in mind, they can easily beadopted to prove an adequate representation of the following units:

CMC unit	*type of data*		*TEI P5 model*
utterance (spoken / multimodal)[7]	transcription of speech	→	<u> (utterance)
bodily activity	textual description	→	<kinesic>
onscreen activity	transcription or textual description	→	<incident>

25 A detailed description of the modelling practices for using <u>, <kinesic> and <incident> for the representation of CMC will be given in Sect. 5. The main challenge from a modelling perspective is the unit *post* as it exhibits both commonalities and differences with spoken utterances and written, monological text so that it cannot be adequately described with any of the models available for the encoding of text and speech in TEI P5.

26 To substantiate this claim we will first characterise some essential properties of posts and then discuss several models from TEI P5 that intuitively come to mind as modelling options before we discuss why these models are not suitable for an adequate representation of posts. The discussion will focus on the properties of *written posts* as the most prominent unit of past and contemporary CMC; *audio posts* will be mentioned where necessary.

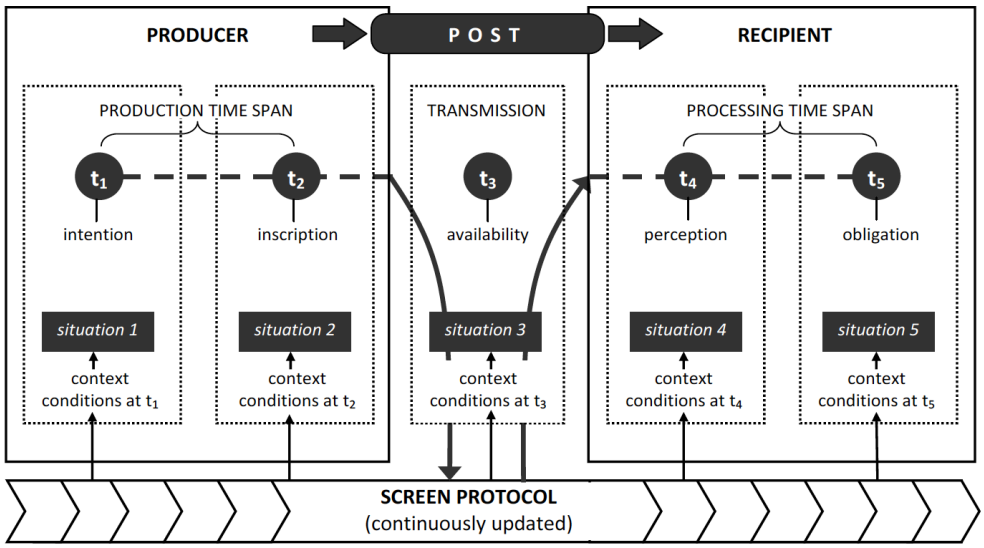
4.1.1 Modelling the *post*: definition and peculiarities

27 A written post is a contribution to an ongoing CMC interaction which (1) has been composed by its author in its entirety as part of a private activity and subsequently (2) has been sent to the server en bloc. Even though in some CMC genres (e.g. WhatsApp) the other parties are informed by an automated alert about the fact that another party is currently composing a new post, they cannot track the process of verbalisation, i.e. how the written utterance emerges in the entry form on the screen

interface of its producer. It is not until the producer performs a 'posting' action (e.g., by hitting the 'enter' key or by activating a 'send' button) that the result of the composition process –the post– is made available for the other parties. From the perspective of its addressees/readers, a post is a piece of text that has been composed in advance. Unlike a speaker turn in spoken conversation, which is volatile and never available for the recipient in its entirety so that it has to be processed at the same time as it is transmitted, the post is presented to the addressees/readers similar to a little text: It is persistent, and all parts of the message are simultaneously present as a visual element filling a certain space on the screen. It can be scanned and read selectively and/or in non-linear order, and its persistence in the screen protocol allows the reader to read it several times, to copy and paste it into their follow-up posts or to link to it in CMC systems that offer hyperlink or citation mechanisms for relating to previous posts (as is the case in WhatsApp, for example). Since posts are visual entities, the point in time at which a post is made available to its addressees is not necessarily the point in time at which it is perceived and processed by them.[8]

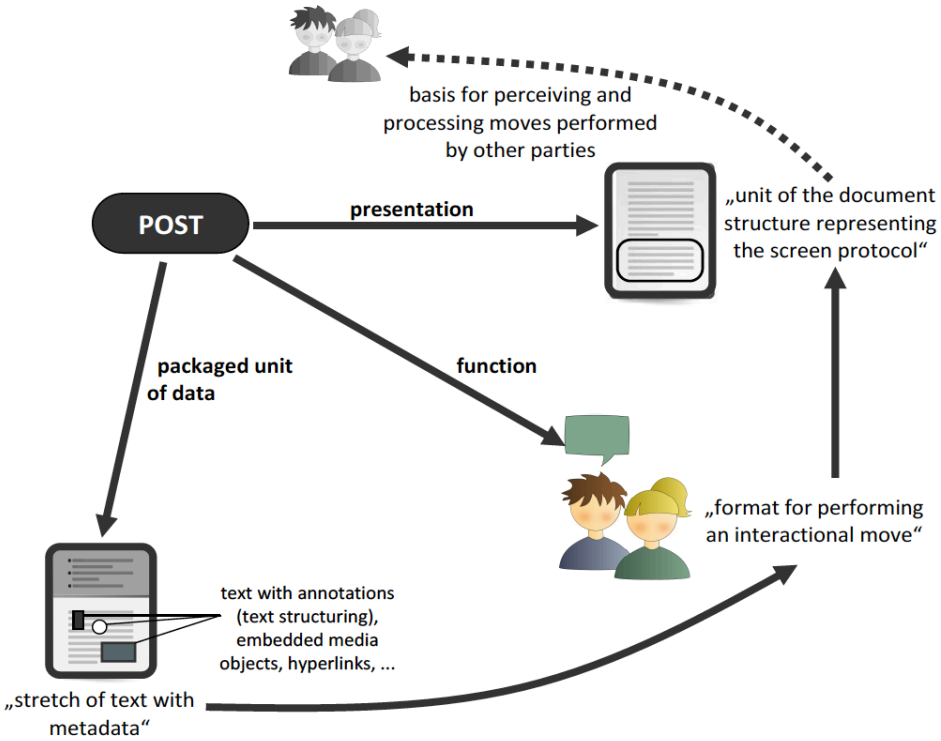
28 Posts are the vehicles through which users of post-based CMC systems can contribute to an ongoing CMC interaction. The format of the post –especially the fact that the process of verbalisation is invisible for the addressees and that the processes of production (producer) and processing (recipient) are dissociated– strongly changes the way how interlocutors can organise their interaction. Whereas forms of interaction which are organised on the basis of turn negotiation necessarily need to rely on simultaneous mutual perception between the involved parties as a resource for interactional management, interaction in post-based CMC –even in 'synchronous' forms like chats– is always non-simultaneous and, therefore, asynchronous. Turn-negotiation is therefore not available as a structuring principle for that type of discourse, and the sequential context of a post can change (several times) between the point in time at which it is planned, it is produced, it is transmitted and it is perceived and processed by an addressee (see the visualisation in Fig. 2). As long as the producer of a post does not check the screen protocol for possible changes of the interactional sequence during production, the sequential context in which the resulting entry (after having submitted the result of production to the server) is embedded in the screen protocol can differ from the context for which it was originally designed. Especially in 'synchronous' forms of CMC, this leads to irregular patterns in the screen protocol which have been termed 'phantom adjacency' (Garcia & Baker Jacobs 1999), 'interleaved exchange sequences' (Herring 1999), 'scrambled interactional patterns' (Storrer 2001: 12) or 'sequential overlaps' (Imo 2015: 23) in previous literature on interactional management in CMC.

Fig. 2. Temporal peculiarities of CMC interaction (instantiation: chat)
(Beißwenger 2016, 2020)



29 From a technology perspective, a post is a chunk of text that is transmitted from a client to a server and then delivered from the server to the clients of all relevant parties. It is treated like a short document which consists of a section with user-generated content and additional metadata (authorID, time stamp, threadID etc.). After its delivery it is represented as a new structural unit in the document which presents and preserves the ongoing interaction on the screens of the interlocutors (screen protocol). From the perspective of the interlocutors, the post is a format for the realisation of contributions to the ongoing interaction (= producers' perspective) and, as a part of the screen protocol, a visual unit that represents a move of a certain other party within the interactional sequence (= recipients' perspective). The different perspectives on the post are visualised in Fig. 3.

Fig. 3. Status and functions of the unit post from the technology and interlocutors' perspective (Beißwenger 2020)



4.1.2 Neither nor <p> nor <u> nor <sp>: Why posts cannot be represented using standard models from TEI P5

30 None of the elements provided in TEI P5 can serve as an adequate model for the representation of posts, even though several elements, at least at first glance, may appear to be a practical solution for the description. In the following we will take a look at possible candidates and discuss why representing posts using these elements may at best be a practical, but in no way a reasonable solution:

31 ○ * (division)* or *<p> (paragraph)*: In view of the fact that the raw data representing logfiles or threads of post-based CMC are typically organised in documents, it may seem obvious to annotate these documents as structured text and treat the posts as a series of *text divisions* or *paragraphs*. The modelling objective, in that case, would be on a description of the surface structure of stored CMC protocols: Similar to other genres of structured text, these protocols consist of divisions and paragraphs, and the boundaries between two instances of posts can be determined using layout information encoded in the original (HTML) markup. Nevertheless, a representation of post-based CMC of that kind would ignore the fact that divisions and paragraphs in text documents typically are the result of a text structure created by an author instance who was responsible for the creation of the text document as a whole. Divisions and paragraphs, in traditional text, are therefore –adopting the macro- and microstructure distinction given in Sect. 3 for traditional text genres– elements of the text *microstructure* which result from decisions of the author instance. Consequently, instances of the elements and <p> in TEI P5 cannot be assigned to different authors. They have been designed as and have proven to be a reasonable solution for the representation of the

structure of monologic text genres; they have not been designed for and could not be adopted in a reasonable way for the representation of dialogic interaction.

32 ○ *<u> (utterance)*: TEI P5 defines utterances as the building blocks of transcriptions of speech. The element <u> is designed to describe a "speech event [which] takes place in time" and is used in spoken conversation to perform a speaker turn (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html, Sect. 8.1). It is a characteristic of speaker turns in conversation that different interlocutors can produce different instances. Similar to CMC documents, a transcript of spoken conversation, on the surface, can be regarded as a structured text document; nevertheless and also similar to CMC documents, a representation *as* structured text would neglect the interactional and dialogic character of that type of discourse and is therefore, as pointed out above, neither an option for the representation of spoken interaction nor of CMC interaction. But why can CMC not be represented adopting the <u> element as a model for the description of posts? This is due to three reasons:

1. Utterances in spoken language unfold under conditions that allow (and force) hearers to process the utterance simultaneously with its production. At the same time, this enables hearers to produce backchannel behaviour parallel to processing the ongoing speaker turn and thus immediately co-work on the further production of the ongoing turn –something which, due to the temporal peculiarities of the post format, is not possible in post-based CMC.
2. The temporal characteristics of the process of producing, processing and responding to posts in post-based CMC much more resembles the process of communicating with monologic texts than the process of producing, processing and responding to utterances in spoken conversation. The production of the post is a composition process which is organised as a private activity and allows for revisions.[9] The processing of a post by an addressee is also a private activity which is invisible for the producer.
3. When producing a written post, the producer can use all sorts of font design and text structuring which are allowed in the post entry form of the CMC system used (capitalisation, different text colour and fonts, bold or italic, paragraph structuring, ...); they may also use hypertext features (e.g. hyperlinking) and embed media objects (images, videos, emojis) so that the resulting product is a digitally stored, multimodal short document. The diverse means of text and hypermedia design applied during the composition of a post cannot be adequately described with the elements available in TEI-P5 for the structuring of spoken utterances. Instead, a representation of the microstructure of post instances requires models which TEI P5 provides for the description of monologic text –even though the unit post as a whole still cannot adequately be described as a division or paragraph (cf. list item 1).

33 ○ *<sp> (speech)*: From our discussion of the TEI P5 models , <p> and <u>, it should have become clear that the representation of CMC posts requires a model which allows for the description of (i) contributions to dialogic sequential exchange which (ii) are realised in a written, text-based mode, may exhibit features of text design and text structuring and appear as structural components of a bigger textual structure (= commonality with or <p>) while at the same time (iii) each of these units may be assigned to a different producer (= commonality with speaker turns/<u>). With the element <sp>, TEI P5 seems to provide a model which combines all of these three characteristics: It is written, it may contain structured text, and each of its instances may be assigned to a speaker. Nevertheless, <sp> is not appropriate for the representation of CMC posts either as it is defined as "an individual speech in a performance text, or a passage presented as such in a prose or verse text" (https://tei-

c.org/release/doc/tei-p5-doc/en/html/ref-sp.html) –a unit whose appearance is restricted to feigned dialogic exchanges occurring in drama texts, verse and prose and, thus, as parts of dialogues which have been produced by one author instance and not naturally evolved between human interlocutors as part of natural occurring interactions. Therefore, <sp> does not seem to be an adequate model for the representation of posts either.

34 As a result from this discussion it seems inevitable to introduce a new element <post> to the TEI encoding framework which is designed (i) to capture both the commonalities and differences of CMC posts with speaker turns and units of structured text and (ii) to present instances of posts as units of naturally evolving, sequential interactions. Our suggestion for a TEI model <post> will be described in Sect. 5.2.

35 Except for the possible occurrence of text design and text structuring features, all characteristics assigned to written CMC posts also hold for audio posts: Audio posts cannot be adequately described as utterances with respect to their temporal peculiarities; they are presented on the screen as visual units and thus as structural (multimodal) components of a bigger, embedding text document; they are transmitted to the addressees at once; they may be edited (in this case: deleted and recorded in a new version) before submitting them to the system; they are persistent and can be listened to several times.

4.1.3 Macrostructure types

36 Besides the introduction of a new element <post>, we need to be able to group sequences of posts to logfiles, threads or other types of CMC macrostructures. This can be achieved without a model extension using the element (for text divisions), considering that a logfile or thread is the biggest structure to be found in the content of a CMC corpus document. Different types of CMC macrostructures can be described as such using the @type attribute which can be assigned parameter values such as “logfile”, “thread” or any other (see the example in Sect. 5.1, Listing 2).

4.2 The microstructure level

37 Based on experiences with modelling CMC data with previous versions of the *CMC-core* schema made in the CLARIN-D and the CoMeRe project, elements on the microlevel of utterances, posts or the textual descriptions of bodily or onscreen activities can be captured using the vast range of models which TEI P5 already provides for the annotation of text structures and elements of spoken language. It is our aim to keep the basic CMC schema provided as *CMC-core* simple and practical so that it can easily be used for a broad range of CMC genres. Nevertheless, we do not preclude that for some corpus projects, depending on how fine-grained the corpus data is planned to be annotated on the microlevel, there may be the wish or need to further adapt some of the elements from TEI P5 to peculiarities of the data. In this case, we recommend to use the *CMC-core* schema and further customise it following the customisation guidelines defined in TEI P5 (https://tei-c.org/guidelines/customization/).

5. *CMC-core* in a nutshell: extensions for representing CMC in TEI P5

38 *CMC-core* describes the basic architecture and models needed for the representation of basic CMC units and the structure of corpus documents that contain written or multimodal CMC discourse. The schema has been designed to meet the six requirements described in Sect. 2. It is based on models from previous schema versions and on modelling experiences from using these schemas for corpus annotation. As a result of previous modelling experiences, and to provide a simple and easy basic schema, it has been created from the previous schema drafts adopting a "reduce to the max" strategy. At present, it has the status of a TEI customisation which is intended to be submitted to the TEI for consideration in a future version of the TEI guidelines. Compared with the current version of the TEI guidelines, this proposal would be minimally invasive while at the same time would allow for a maximum scope of applications concerning the representation and use of CMC resources and corpora in the humanities context and beyond.

39 In the following, we describe the additions to the TEI that we defined in our customisation *CMC-core*, using the technical vocabulary of the TEI infrastructure and giving listings of TEI encodings of CMC phenomena from different genres as an illustration. The customisation has been published as an ODD ("one document does it all"), which is a TEI document that formally defines and documents the customisation. In our examples, we used numerous encoding features of the TEI which are already available, such as <figure> for emojis, <time> for timestamps, or @synch, @start and @end for references to a timeline. However, we do not discuss these here because we focus on the new features introduced in the *CMC-core* customisation.

40 Below is a short overview of the four types of specifications our ODD for *CMC-core* introduces, elaborated in the subsequent sections:

(1) A new **module** named *cmc* is introduced. It is referenced by the new model class *model.divPart.cmc*, by the new attribute class *att.global.cmc*, and by the new element <post> (Sect. 5.1).
(2) The new **model class** *model.divPart.cmc* is introduced. It is defined to be a member of the existing class *model.divPart* and serves as a container for the new element <post> (Sect. 5.2).
(3) The new **element** <post> is introduced along with several attributes (Sect. 5.3).
(4) The new **attribute class** *att.global.cmc* is introduced. It defines the new attribute @creation to be available for all elements. The existing attribute class *att.global* is defined to additionally be a member of the new class (Sect. 5.4).

5.1 The module *cmc*

41 A module in the TEI is a thematically defined group of specifications of elements, attributes and classes. In TEI P5, there are 21 pre-defined modules such the ones for *textstructure*, *corpora*, or *performance texts* (TEI P5, 2007). One purpose of a module is to group genre-specific encoding features so that all of them are included or excluded at once when the module is selected or deselected in a customisation. We hereby propose a new, separate module "cmc" to group the newly proposed (and future) CMC-specific features. We think that, conceptually, the introduction of a separate CMC module is justified. However, we are aware that the introduction of a new module constitutes a

fairly big step for an extension of the TEI and we would alternatively also be satisfied if our proposed models could be included in the official TEI in existing modules.

5.2 The model *model.divPart.cmc*: CMC macrostructures

42 Models serve to define the allowed content of markup elements in the TEI. Models can be combined to form more complex models, thus forming a membership hierarchy. The *CMC-core* model *model.divPart.cmc* introduces the new element <post> and is defined to be a member of *model.divPart*, hence making <post> available on the divPart level, i.e. as possible content of a TEI (division) element. This allows for using and combining occurrences of <post>, <u>, <kinesic>, <incident> (and further elements) within one and the same , or directly within a <body>, in order to represent the interleaved occurrences of posts, utterances (TEI <u>) and non-verbal acts i.e. bodily or onscreen activities (TEI <kinesic> or <incident>) as e.g. on the GUI (graphical user interface) of a multimodal CMC environment. Listing 1 shows how a spoken utterance <u>, an avatar's bodily activity <kinesic>, and a written post <post> occur on the same level within the text <body>, representing parts of a multimodal chat in Second Life from the Archi21 corpus. Listing 2 shows a sequence of <post> instances within one , representing a discussion thread on a Wikipedia talk page.

Listing 1. Second Life multimodal chat example, adapted to *CMC-core*, from Chanier & Wigham (2015)

```
<text>
  <body>
    <u xml:id="cmr-archi21-slrefl-es-j3-1-a191" who="#tingrabu"
       start="#cmr-archi21-slrefl-es-j3-1-ts373"
       end="#cmr-archi21-slrefl-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 euh we have not
       time and hm <pause dur="PT1S" /> too quickly sorry
      […]
    </u>
    <kinesic xml:id="cmr-archi21-slrefl-es-j3-1-a192" who="#romeorez"
       start="#cmr-archi21-slrefl-es-j3-1-ts376"
       end="#cmr-archi21-slrefl-es-j3-1-ts377" type="body"
       subtype="kinesics">
       <desc>
          <code>eat(popcorn)</code>
       </desc>
    </kinesic>
    […]
    <post mode="written" creation="human"
       xml:id="cmr-archi21-slrefl-es-j3-1-a195" who="#tfrez2"
       start="#cmr-archi21-slrefl-es-j3-1-ts380"
       end="#cmr-archi21-slrefl-es-j3-1-ts381" type="chat-message">
       <p>it went too quickly?</p>
    </post>
    [...]
```

Listing 2. Discussion thread on a Wikipedia talk page. The encoding of Wikipedia talk uses a further customisation in which <signed> is allowed to occur inside <p>, which is, however, not part of *CMC-core* (cf. Sect. 4.2)

```
<div type="thread">
  <head>Naturally occurring?</head>
  <post mode="written" xml:id="p4" indentLevel="0" who="#u005"
    synch="#t005">
    <p>I'm not sure that this is a proper criterium, 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="#t006">
    <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>
```

5.3 The element <post> and its attributes: Basic unit of CMC encoding

43 A post is defined in the ODD gloss as a written (sometimes spoken) contribution to an ongoing CMC interaction which (1) has been composed by its author in its entirety as part of a private activity and (2) has been sent to the server en bloc. It is defined to be a member of *model.divPart.cmc* (see Sect. 5.2) i.e. it can co-occur together with <u>, <kinesic>, <incident> and other existing TEI elements within a (text division) or directly within the <body> (text body).

44 The content model of <post> is based on the existing *macro.paraContent*, which is a TEI macro for the content model of <p> and <u>. We added *model.headLike, model.pLike*, the element <opener> and *model.divBottom* in order to include the possibility of headings, paragraphs, openers, closers and salutations within <posts>.

45 Posts are defined so that they can take all sorts of attributes, i.e. <post> is defined to be a member of the available TEI attribute classes *att.ascribed, att.canonical, att.datable, att.global, att.timed*, and *att.typed*. In the CMC corpora provided by the members of the TEI CMC SIG, uses of the TEI attributes @who, @synch, @type, @subtype, @rend, and @xml:id for <post> abound.

46 Additionally, three new attributes are defined specifically for <post> in *CMC-core*:

(1) @mode encodes the basic distinction between written and spoken posts. It may take one of the two values “written” and “spoken” (cf. the encoding of a sequence of a written and spoken post in a WhatsApp chat interaction in Listing 3).
(2) @replyTo serves to indicate to which previous post the current post replies or refers to. This attribute should be used to encode “technical reply” information that is part of the original metadata of the post due to a formal reply action by the

user in the CMC environment (such as activating a reply button). It should rather not be used to encode interpreted or inferred reply relations based on linguistic cues or discourse markers, nor for the relations expressed by indentations on wiki talk pages (cf. Lüngen & Herzberg 2019). Its TEI value type is data.pointer i.e. it contains a list of IDs of previous posts (cf. the encoding of a blog comment replying to a previous comment in Listing 4).
(3) @indentLevel marks the level of indentation of the current post in a thread-like structure (as defined by its author and in relation to the standard level of non-indentation which is to be encoded as @indentLevel="0"). It is used in wiki talk corpora but may also be used for genres such as weblog comments when an HTML encoding was used as a source. Its TEI value type is data.count (cf. the encoding of the Wikipedia discussion thread in Listing 2).

47 A further newly defined attribute, @creation, can be seen in our encoding examples, which is however not specific to the <post> element and described in Sect. 5.4 below.

Listing 4. Written and spoken post in WhatsApp chat interaction including an emoji, adapted to CMC-core. From the corpus MoCoDa2 (2018)

```
<post mode="spoken" creation="human" synch="#t003" who="#A05"
    xml:id="m7"> Sagt Anne auch gerade. JA! Kann ich zustimmen. </post>
<post mode="written" creation="human" synch="#t003" 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 3. A blog comment, replying to a previous comment. From the Scilogs corpus, adapted to *CMC-core* (Grumt Suárez et al. 2016)

```
<post xml:id="p5" type="comment" who="#u4" synch="#t005" replyTo="#p4">
   <p>“Wenn Sie diesen Gruppen also “mangelnde Bildung“ attestieren  wollen,
      so verwenden Sie bereits einen bestimmten,  kulturgebundenen Bildungsbe
      griff.”</p>
   <p>Ich hoffe doch, wir können beim Bildungsbegriff der Aufklärung
      bleiben. Wer das nicht möchte, hat die Wissenschaft verlassen.</p>
</post>
```

Listing 5. Twitter interaction including a retweet. This sequence of tweets is taken from the view of one user's (u1) timeline. The retweet and the retweeted tweet are encoded as two separate tweets (p2 and p3) with their own sets of metadata and linked by the reference to p3 in the @ref attribute of tweet p2. The retweet p2 is an otherwise empty <post> (however retweets may also have some content of their own)

```
<post mode="written" key="1043764753502486528" type="tweet"
    creation="human" synch="#tweetsbcrn18.t001" xml:id="p1" who="#u1"
    xml:lang="deu">
    <time creation="system"> 12:31 </time> Heute mit super
    Unterstützung, wir grunzen, wenn die Zeit vorbei ist. <ref
    type="hashtag" target="https://twitter.com/hashtag/bcrn18?src=hash"
     >#bcrn18</ref>
    <ref type="hashtag" target="https://twitter.com/hashtag/wikidach?src=hash"
     >#wikidach</ref> PS: Die beiden brauchen noch Namen. Hinweise dazu am
     Empfang abgeben! <ref type="twitter-account"
     target="https://twitter.com/AndreLo79">@AndreLo79</ref>
    <figure type="image" creation="human">
        <graphic
            url="https://pbs.twimg.com/media/DnwygdSW4AAoTUn.jpg:large"/>
    </figure>
</post>
<post mode="written" key="1043769240136880128" creation="unspecified"
     type="tweet" subtype="retweet"  who="#u1" ref="#p3"
     synch="#tweetsbcrn18.t002" xml:id="p2" />
<post mode="written" key="1043767827927388160" creation="human"
     type="tweet" who="#u3" synch="#tweetsbcrn18.t002" xml:lang="de"
     xml:id="p3">
   <time creation="system"> 12:43 </time>
   <figure type="image" creation="human">
      <graphic
          url="https://pbs.twimg.com/media/Dnw1TRNXgAAKqlK.jpg:large"/>
   </figure>
</post>
```

Listing 6. Chat system message: user enters a chatroom, "system" listed as a participant in particDesc//listPerson, example from the Dortmund Chat Corpus 2.2 (2016)

```
<listPerson>
    <person xml:id="SYSTEM">
        <persName>system</persName>
    </person>
    <!-- [...] -->
</listPerson>
<!--  [...] -->
<post type="event" creation="system" who="#SYSTEM" rend="color:navy">
  <p>
    <name type="nickname" corresp="#A07">Interseb</name> betritt den Raum.</p>
</post>
```

5.4 The attribute class *att.global.cmc*: The global attribute @creation

48 The new global attribute @creation, introduced in the new attribute class *att.global.cmc*, may indicate *for any TEI element* how its content was created in a CMC environment i.e. whether it was created by a human user, whether it was created by the respective CMC system (e.g. a status message, a timestamp), or whether the user activated a template that subsequently generated the textual content, such as in a signature. The attribute is optional since it might not be appropriate to use it when all text content has been

created in the same mode and no distinction needs to be made. Based on our corpora, we suggest a closed list of five possible values.

1. “human”: when then content of the respective element was “naturally” typed or spoken by a human user (cf. the chat posts in Listing 3)
2. “template”: when the content of the respective element was generated after a human user activated a template for its insertion (cf. <signed> i.e. signatures in wiki talk in Listing 2)
3. “system”: when the content of the respective element was generated by the system, i.e. the CMC environment (cf. the system message in an IRC chat in Listing 6)
4. “bot”: when the content of the respective element was generated by a bot i.e. a non-human agent, mostly external to the CMC environment (no example)
5. “unspecified”: when it is unspecified or unknown how the content of the respective element was generated (cf. the <post> of the retweet p2 in Listing 5)

6. Outlook

49 The field of computer-mediated communication is mature enough to be considered for coverage in a future version of the TEI guidelines. With a minimal number of additions to the existing framework, the TEI could provide a model to ensure interoperability in the growing field of resources and applications that are dedicated to the collection, processing and analysis of data from CMC genres. Being one of the success stories in the field of digital humanities, a “CMCified” version of the TEI could reach out even beyond the humanities to the vast inter- and multidisciplinary field of social media analysis.

50 The *CMC-core* customisation described in this article is scheduled to be submitted as a request to the TEI Technical Council in late 2019. Its ODD file, together with sample encodings including the full TEI documents containing the examples given in Sect. 5, can be retrieved from the TEI wiki at https://wiki.tei-c.org/index.php?title=SIG:CMC. A derived RNG schema which is also available at this location can immediately be used for corpus annotation.

BIBLIOGRAPHY

Corpora

Chanier, T. & Wigham, C.R. (2015). Archi21 corpus: collaborative language and architectural learning in Second Life. Banque de corpus CoMeRe. Ortolang.fr: Nancy. https://hdl.handle.net/11403/comere/cmr-archi21/cmr-archi21-tei-v1 (last visited 2019-08-20).

CoMeRe (2015): https://repository.ortolang.fr/api/content/comere/v2/comere.html (last visited 2019-08-20).

Dortmund Chat Corpus 2.2 (2016): http://hdl.handle.net/10932/00-03B0-14FA-A8D0-0F01-F (last visited 2019-08-20).

Reffay, C. Chanier, T. Lamy, M.-N. & Betbeder, M.-L. (2009). (editors). LETEC corpus Simuligne [corpus]. MULCE.org: Clermont Université. http://repository.mulce.org (last visited 2019-08-20).

Wikipedia Corpora at IDS (2014ff): http://www1.ids-mannheim.de/kl/projekte/korpora/archiv/wp.html (last visited 2019-08-20).

MoCoDa2 (2018): https://db.mocoda2.de (last visited 2019-08-20).

Bibliography

Beißwenger M. (2007). *Sprachhandlungskoordination in der Chat-Kommunikation.* (Linguistik – Impulse & Tendenzen 26). Berlin/New York: de Gruyter.

Beißwenger M. (2010). "Chattern unter die Finger geschaut: Formulieren und Revidieren bei der schriftlichen Verbalisierung in synchroner internetbasierter Kommunikation", *in* V. Ágel & M. Hennig (ed.) *Nähe und Distanz im Kontext variationslinguistischer Forschung.* (Linguistik – Impulse & Tendenzen 35). Berlin/New York: de Gruyter, 247–294.

Beißwenger M. (2016). "Praktiken in der internetbasierten Kommunikation", *in* A. Deppermann, H. Feilke & A. Linke (ed.) *Sprachliche und kommunikative Praktiken.* Jahrbuch des Instituts für Deutsche Sprache 2015. Berlin/Boston: de Gruyter, 279–310.

Beißwenger M. (2018). "Internetbasierte Kommunikation und Korpuslinguistik: Repräsentation basaler Interaktionsformate in TEI", *in* H. Lobin, R. Schneider & A. Witt (ed.) *Digitale Infrastrukturen für die germanistische Forschung*. (Germanistische Sprachwissenschaft um 2020 6). Berlin/Boston: de Gruyter, 307–349.

Beißwenger M. (2020, in press). "Internetbasierte Kommunikation als Textformen-basierte Interaktion: ein neuer Vorschlag zu einem alten Problem", *in* H. Lobin, K. Marx & A. Schmidt (ed.) *Deutsch in sozialen Medien: interaktiv, multimodal, vielfältig*. Jahrbuch 2019 des Leibniz-Instituts für Deutsche Sprache. Berlin/New York: de Gruyter.

Beißwenger M., Chanier T., Erjavec T., Fišer D., Herold A., Lubešic N., Lüngen H., Poudat C., Stemle E., Storrer A. & Wigham C. (2017). "Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries", *Selected Papers from the CLARIN Annual Conference 2016, October 26–28 2016*. France, Aix-en-Provence: Linköping University Electronic Conference Proceedings. http://www.ep.liu.se/ecp/contents.asp?issue=136 (last visited 2019-08-20).

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* 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476) (last visited 2019-08-20).

Beißwenger M., Imo W., Fladrich M. & Ziegler E. (2019). "*https://www.mocoda2.de:* a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions", *European Journal of Applied Linguistics* 7 (2).

Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C., Hriba L., Longhi J. & Seddah D. (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 (last visited 2019-08-20).

Garcia A. C., Baker Jacobs J. (1999). "The Eyes of the Beholder: Understanding the Turn-Taking System in Quasi-Synchronous Computer-Mediated Communication", *Research on Language and Social Interaction* 32 (4): 337–367.

Grumt Suárez H., Karlova-Bourbonus, N. & Lobin, H. (2016). “Compilation and Annotation of the Discourse-structured Blog Corpus for German”, *in* Proc. 4th Conference on CMC and Social Media Corpora for the Humanities. Ljubljana: University of Ljubljana, 26–29. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-conference-proceedings-2016.pdf (last visited 2019-08-20).

Herring S. C. (1999). “Interactional Coherence in CMC”, *Journal of Computer-Mediated Communication* 4 (4). http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1999.tb00106.x/full (last visited 2019-08-20).

Imo W. (2015). *Vom Happen zum Häppchen … Die Präferenz für inkrementelle Äußerungsproduktion in internetbasierten Messengerdiensten.* (Networx 69). https://www.mediensprache.net/networx/networx-69.pdf (last visited 2019-08-20).

Jannidis F. (2017). “Grundlagen der Datenmodellierung”, *in* F. Jannidis, H. Kohle & M. Rehbein (ed.) *Digital Humanities. Eine Einführung.* Stuttgart: J. B. Metzler, 99–108.

Lobin H. (2010). *Computerlinguistik und Texttechnologie*. (= Linguistik für Bachelor 3282). Paderborn: Wilhelm Fink.

Lüngen H., Beißwenger M., Herold A. & Storrer A. (2016). “Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN”, *in* S. Dipper, F. Neubarth & H. Zinsmeister (ed.) *Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), 1561–64.

Lüngen, H. & Herzberg L. (2019). “Types and annotation of reply relations in computer-mediated communication”, *European Journal of Applied Linguistics* 7 (2). https://doi.org/10.1515/eujal-2019-0004 (last visited 2019-08-20).

Lüngen H. & Kupietz M. (2017). “CMC Corpora in DeReKo”, *in* P. Bański, M. Kupietz, H. Lüngen, P. Rayson, H. Biber, E. Breiteneder, S. Clematide, J. Mariani, M. Stevenson & T. Sick (ed.) *Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017*. Mannheim : Institut für Deutsche Sprache, 20–24.

Margaretha E. & Lüngen H. (2014). “Building Linguistic Corpora from Wikipedia Articles and Discussions”, *Journal of language Technology and Computational Linguistics* 29 (2): 59–82. https://jlcl.org/content/2-allissues/6-Heft2-2014/3MargarethaLuengen.pdf (last visited 2019-08-20).

Schröck J. & Lüngen, H. (2015). “Building and Annotating a Corpus of German-Language Newsgroups”, *in* M. Beißwenger & T. Zesch (ed.) *NLP4CMC 2015. 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media. Proceedings of the Workshop, September 29, 2015 University of Duisburg-Essen, Campus Essen.* German Society for Computational Linguistics & Language Technology (GSCL), 17–22.

Sperberg-McQueen C. M. (2018). “Kernideen der deskriptiven Textauszeichnung”, *in* H. Lobin, R. Schneider & A. Witt (eds.) *Digitale Infrastrukturen für die germanistische Forschung.* (Germanistische Sprachwissenschaft um 2020 6). Berlin/Boston: de Gruyter, 292–305.

Storrer A. (2001). “Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation”, *in* M. Beißwenger (ed.) *Chat-Kommunikation. Sprache, Interaktion, Sozialität und Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* Stuttgart: Ibidem-Verlag, 3–24.

[TEI P5] TEI Consortium (2007*). TEI P5: Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/Guidelines/P5/ (last visited 2019-08-20).

Wigham C. R. & Chanier T. (2013). “Interactions between text chat and audio modalities for L2 communication and feedback in the synthetic world Second Life”, *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2013.851702 (last visited 2019-08-20).

Wigham, C. R. & Ledegen G. (ed.) (2017). *Corpus de Communication Médiée par les Réseaux*. Paris: L’Harmattan.

NOTES

1. *Communication Médiée par les Réseaux* – French for CMC.
2. We apply the distinction between CMC *modes* and *modalities* following Chanier et al. (2014: 6) according to whom the term *mode* is used to characterise certain types of semiotic resources while *modality* refers to specific ways of realising communication (i.e. CMC technologies).
3. We are aware of the fact that in certain audio and video conferencing systems it is possible to record the whole session. Nevertheless, a recording of an audio or video conference –in contrast to a chat logfile– cannot be used during the session as a documentation of previous speaker turns in order to keep track of the ongoing interaction. Instead, it can only be used after the session as a documentation of the finished interaction as a whole.
4. The restriction *near-* in the term ‚near-simultaneous’ takes into account that in audio and video conferences there may be little delays between speaker behavior (spoken, mimic, gesture) and its transmission, or that, when interacting using a virtual avatar in a 3D environment such as *Second Life,* the user first has to enter the command for performing a certain kinesic move before the avatar performs the respective move on screen.
5. We mention this type of CMC unit for the sake of completeness, but it will not be considered as a *model* in the *CMC-core* TEI schema we describe in Sect. 5.
6. A detailed description and substantiation of the *post* model is given in Beißwenger (2016, 2018, 2020).
7. As mentioned in Sect. 3, our schema only covers spoken utterances. Even though written utterances did occur in a few CMC systems (*UNIX Talk*), they do not play a role in contemporary CMC anymore and therefore are ignored in *CMC-core*.
8. Empirical evidence for the temporal shift between production, availability and perception of posts in chat interactions is given in Beißwenger (2007, 2010).
9. This holds for written posts as well as for audio posts: The production of an audio post can be cancelled before transmission and started anew if the producer is not satisfied with it without the addressees taking notice of the message produced so far –something which is impossible under the conditions of spoken conversations.

ABSTRACTS

In this Paper, we describe a schema and models which have been developed for the representation of corpora of computer-mediated communicatin (CMC corpora) using the representation framework provided by the Text Encoding Initiative (TEI). We characterise CMC discourse as dialogic, sequentially organised interchange between humans and point out that many features of CMC are not adequately handled by current corpus encoding schemas and tools.

We formulate desiderata for a representation of CMC in encoding schemes and argue why the TEI is a suitable framework for the encoding of CMC corpora. We propose a model of basic CMC units (utterances, posts, and nonverbal activities) and the macro- and micro-level structures of interactions in CMC environments. Based on these models, we introduce CMC-core, a TEI customisation for the encoding of CMC corpora, which defines CMC-specific encoding features on the four levels of elements, model classes, attribute classes, and modules of the TEI infrastructure. The description of our customisation is illustrated by encoding examples from corpora by researchers of the TEI SIG CMC, representing a variety of CMC genres, i.e. chat, wiki talk, twitter, blog, and Second Life interactions. The material described, i.e. schemata, encoding examples, and documentation, is available from the of the TEI CMC SIG Wiki and will accompany a feature request to the TEI council in late 2019.

Dans cet article, nous décrivons un schéma et des modèles de représentation développés pour structurer les corpus de communication médiée par ordinateur (CMC) en suivant les recommandations de la Text Encoding Initiative (TEI). Nous considérons le discours CMC comme un échange dialogique entre humains, organisé de manière séquentielle. Nous insistons d'abord sur le fait que de nombreuses caractéristiques de la CMC ne sont pas traitées de manière adéquate par les schémas et les outils actuels d'encodage de corpus. Nous formulons donc un ensemble de recommandations pour représenter la CMC avec des schémas d'encodage, en insistant sur le fait que la TEI nous semble être un cadre particulièrement approprié pour l'encodage des corpus CMC. Nous proposons une modélisation des unités de base de la CMC (énoncés, messages et actions non verbales) ainsi que des structures de niveaux macro- et micro des interactions dans les environnements de la CMC. À partir de ces modèles, nous introduisons le *CMC-core*, un noyau TEI pour l'encodage des corpus CMC, qui définit un ensemble de traits d'encodage spécifiques à la CMC sur quatre niveaux: (i) éléments, (ii) classes de modèles, (iii) classes d'attributs et (iv) modules de l'infrastructure TEI. La description du noyau proposé est illustrée au moyen d'exemples extraits des corpus des chercheurs du groupe SIG TEI CMC, représentant une grande variété de genres de la CMC (le chat, le wiki talk, le tweet, le blog, les interactions Second Life...). Le matériel décrit, *i.e.* les schémas, les exemples d'encodage et la documentation, est disponible sur le Wiki du SIG CMC TEI et accompagnera une demande d'enrichissement de la TEI (*TEI feature request*) au conseil de la TEI à la fin de l'année 2019.

INDEX

Keywords: CMC, interactional linguistics, cmc corpora, standards, TEI
Mots-clés: CMC, linguistique interactionnelle, corpus CMC, standards, TEI

AUTHORS

MICHAEL BEISSWENGER

University of Duisburg-Essen

HARALD LÜNGEN

Leibniz Institute for the German Language, Mannheim