Abstract

The paper presents an XML schema for the representation of genres of computer-mediated communication (CMC) which is compliant with the encoding framework defined by the TEI. It was designed for the annotation of CMC documents in the project “Deutsches Referenzkorpus zur internetbasierten Kommunikation” (DeRiK), which aims at building a corpus on language use in the most popular CMC genres on the German-speaking internet. The focus of the schema is on those CMC genres which are written and dialogic—such as forums, bulletin boards, chats, instant messaging, wiki and weblog discussions, microblogging on Twitter, and conversation on “social network” sites.

The schema provides a representation format for the main structural features of CMC discourse as well as elements for the annotation of units which are often regarded as “typical” for language use on the internet. The schema introduces an element posting which describes the stretches of text that are sent to the server by one user at a certain point in time. Postings are the main constituting elements of threads and logfiles which, in our schema, are described as the two main types of CMC macrostructures. For the microlevel of CMC documents (= the structure of the posting content), the schema introduces elements for selected features of “internet jargon” such as emoticons, interaction words and addressing terms. It allows for an easy anonymization of CMC data for purposes in which the annotated data shall be made publicly available and includes metadata which are necessary for referencing random excerpts from the data as references in dictionary entries or as results of corpus queries.

A documentation of the schema as well as encoding examples can be retrieved from the web at http://www.empirikom.net/bin/view/Themen/CmcTEI. The schema is meant to be a core model for representing CMC which can be modified and extended by others according to their own specific perspectives on CMC data. It could be a first step towards an integration of features for the representation of CMC genres into a prospective new version of the TEI Guidelines.

Keywords: computer-mediated communication, cmc, web genres, thread, logfile, forum, chat
1 Introduction

In the past three decades, computer networks and especially the internet have brought forth new and emerging genres of interpersonal communication (“computer-mediated communication”, henceforth *CMC*)—such as the e-mail, online forums, chats, instant messaging, or weblogs. In several respects, these genres stand in the tradition of well-known genres such as spoken conversation or written letters. On the other hand, they display linguistic and structural features which differ both from speech and from written text and which can be traced back to the impact of their technological frameworks as well as to the ways in which interlocutors adapt to their potentials and limitations.

Recent surveys on the use of the internet (such as, e.g., the annually conducted German “ARD/ZDF-Onlinestudie”) show that the use of CMC applications makes up an important part of everyday communication. To get to a better understanding of these new forms of mediated communication and their linguistic peculiarities, we need tools and models that allow one to analyze them on a broad empirical basis and with the help of corpus technology and methods from computational linguistics. One important prerequisite for that would be a common format for the representation and exchange of CMC resources. Even though CMC...
phenomena are not a completely new field of research within the humanities anymore, such a format still does not exist.

In this paper, we present an XML schema for the representation of genres of computer-mediated communication which is conformant with the encoding framework defined by the TEI. Up to now, CMC genres and document types have not yet been in the focus of the TEI. Therefore our schema takes the modules as well as the element and attribute classes of the P5 version of the TEI Guidelines (released on November 1, 2007) as a starting point and uses the TEI customization mechanism in order to adapt their use for an area of resources that is not yet covered by them.

The focus of the schema is on those CMC genres which are written and dialogic—forums, bulletin boards, chats, instant messaging, wiki and weblog discussions, microblogging on Twitter, and conversation on “social network” sites. The schema has been developed in the context of the project “Deutsches Referenzkorpus zur internetbasierten Kommunikation” (DeRiK), which is a joint initiative of TU Dortmund University and the Berlin-Brandenburg Academy of Sciences and the Humanities (BBAW). The project is embedded in the scientific network “Empirische Erforschung internetbasierter Kommunikation” (http://www.empirikom.net/), funded by the Deutsche Forschungsgemeinschaft (DFG). The aim of the project is to build a corpus on language use on the German-speaking internet which covers the most popular CMC genres. The corpus is designed to be integrated into the corpora and lexical resource framework provided by the project “Digitales Wörterbuch der deutschen Sprache” (DWDS, http://www.dwds.de) at the BBAW “Zentrum Sprache”.

Since all corpus resources of the DWDS project are already encoded according to the TEI encoding framework and since up to now there is no common standard for an XML/TEI representation of the structural and linguistic properties of CMC resources, the project group decided that the TEI standards would be an optimal basis for the annotation of the DeRiK data—assuming that the encoding framework of the TEI proves to be flexible enough to be adapted to the particularities of CMC discourse. In particular, we formulated the following requirements for our schema:

- It should provide a model that is adapted to the structural particularities of CMC discourse and that takes into consideration that the interlocutors’ contributions to conversations in forums, chats, in wiki and weblog discussions, etc. can neither be adequately described as utterances in speech nor as paragraphs in traditional writing;
- it should provide elements for the annotation of units which are often regarded as “typical” for language use on the web and which are of special interest for everybody who wants to compare linguistic features of CMC discourse with the language documented in text corpora (such as the DWDS corpora); in the DeRiK context, a special focus lies on units which we subsume under the category interaction signs which includes, amongst others, emoticons, interaction words, and addressing terms;
- it should be open for extensions by other researchers in the field of empirical CMC research or by corpus designers who want to adapt the schema for their own project purposes (especially on the microlevel, which—in the terminology of our project—is the level below the individual user contribution);
- on the macrolevel (= the level above the individual user contributions), its structure should be oriented on surface phenomena and, thus, be as independent as possible from any specific theory of CMC discourse; this will allow use of the macrostructure model

---

1 For a brief description of the project, see http://www.empirikom.net/bin/view/Themen/DeRiK.
of the schema as a basic document structure in as many projects as possible; in addition, it will allow automatization of the generation of the basic TEI structure of CMC documents (which is an important requirement, especially in projects that aim at building large corpora);

- it shall allow for an easy (but reversible) anonymization of CMC data for purposes in which the annotated data should be made available as a resource for other researchers or for the public (as is intended with the DeRiK corpus as part of the DWDS framework);
- it shall provide all information and metadata which are necessary for using and referencing random excerpts from the data as references in a general language dictionary as well as in the results of a corpus query (as is the case in the DWDS online portal at http://www.dwds.de).

In the following (section 2), we will first give an outline of the motivation and project context which form the background or our work. We then will describe the design of our schema in detail and illustrate some of our basic modeling decisions with the help of examples from our data (section 3). The schema itself, its documentation, and some encoded example documents can be retrieved from the web at http://www.empirikom.net/bin/view/Themen/CmcTEI.

The current version of the schema will form the basis for the annotation of CMC documents in the DeRiK context. Since it is meant to be a core model for representing CMC, it can be modified and extended by others according to their own specific perspectives on CMC data. It will have to prove its adequacy for the resource types in focus by being used and analyzed by more researchers and corpus builders than just its authors. The schema and its further discussion could be a first step towards an integration of features for the representation of CMC genres into a prospective new version of the TEI Guidelines.

2 Motivation and Project Background

2.1 Motivation

The motivation for building a corpus of German CMC is to close a gap in the field of corpora which are currently available for CMC as well as for contemporary German in general: In the area of computer-mediated communication up to now hardly any annotated specialized corpora exist. Also the general corpora of contemporary German do not systematically include the language use on the internet. This poses a blatant gap, since over the past years online communication has become an important part of everyday communication and, thus, can no longer be left out when documenting contemporary everyday language use. Corpus linguistics is aware of that gap. In addition to the DeRiK project which aims at building a German CMC corpus and integrating it into the DWDS general language corpora, there are similar ideas or projects for other languages as well. One example is the SoNaR project which aims at building a balanced reference corpus of contemporary Dutch in which an own subcorpus of CMC shall be included (Reynaert et al. 2010).

Due to a lack of standards for representing CMC, up to now corpus-based research projects on features of CMC discourse have typically developed their own, project-specific standards.

---

2 We would like to thank the members of the scientific network Empirikom as well as Laurent Romary and the participants of the Annual Conference and Members’ Meeting of the TEI Consortium 2011 in Würzburg for valuable discussions on the subject and for their comments on previous versions of the schema.
encoding schemas (see, e.g., the XML encoding for chats that has been designed for the resources included in the Dortmund Chat Corpus, 2003-2009\(^3\)). This complicates, if not even inhibits, the sharing of the data across projects. This is all the more regrettable because the individual projects add valuable structural and semantic information to their data through their annotations (not to mention the time and manpower that it takes for annotating the data). The potentials of sharing, merging and comparing corpora, particularly in contrastive research, call for a basic schema which suits the needs of various projects and which is easy to handle and extend.

In addition, such a schema should be compliant with encoding frameworks which are already widely used in existing text and speech corpora. This would allow the schema to not only meet the needs of scholars who are interested in just CMC but also the needs of all scholars who are interested in phenomena of contemporary language in general or in doing comparative analyses of linguistic phenomena in CMC corpora and in corpora of “traditional” text or speech genres.

Since many resources within the humanities are already using the encoding framework provided by the Text Encoding Initiative (TEI), a basic schema for CMC would ideally comply with the TEI format. As will be shown in section 3 of this paper, the TEI format has the power and flexibility to describe CMC structures and features even though modules or elements which cover the particularities of CMC discourse are not yet implemented in the TEI. Therefore, a TEI-compliant XML schema for CMC discourse, right now, can only be designed using customizations. Considering the relevance of the internet as a communication medium, an own module for CMC document types and features could be an important extension for prospective follow-up versions of the TEI Guidelines.

### 2.2 The DeRiK Corpus in the Context of the DWDS System

Designers of balanced corpora representing the current state of a language, e.g. German, should be sure to include all relevant types of genres in which the contemporary use of this language is embodied. Nowadays, this should include genres of computer-mediated communication. In the project “Deutsches Referenzkorpus zur internetbasierten Kommunikation” (DeRiK), we are aiming at building a corpus of German CMC which covers data from the most popular CMC genres\(^4\). The sampling of the data is guided by the findings of the “ARD/ZDF-Onlinestudie”, a German online usage survey conducted annually (http://www.ard-zdf-onlinestudie.de) which shows which genres are most frequently used by German online users. For practical reasons, though, the project will set out with sampling such domains and genres that are cleared from intellectual property rights.

DWDS (“Digitales Wörterbuch der deutschen Sprache”, http://www.dwds.de) is a digital lexical system developed by and hosted at the BBAW. The system offers one-click-access to three different types of resources (Geyken 2007):

- a) lexical resources: a common language dictionary\(^5\), an etymological dictionary, and a thesaurus;

---

\(^3\) http://www.chatkorpus.tu-dortmund.de

\(^4\) http://www.empirikom.net/bin/view/Themen/DeRiK

\(^5\) This dictionary is based on a six-volume printed dictionary, the “Wörterbuch der deutschen Gegenwartssprache” (WDG, en.: “Dictionary of Contemporary German”) published between 1962 and 1977 and compiled at the Deutsche Akademie der Wissenschaften (cf. [WDG]).
b) corpus resources: a balanced reference corpus (called “DWDS core corpus”) of German ranging from 1900 up to now. The corpus is balanced wrt. a broad typology of texts. It contains nearly equal shares of journalistic texts, scientific prose, functional texts, and fiction. Up to now, CMC did not play a role, neither as an independent text type nor as a part of one or more of these text types;

c) a set of additional newspaper corpora and specialized corpora (e.g. German newspapers of Jewish communities edited in the first decades of the 20th century);

d) statistical resources for words and word combinations.

On the web frontend, these resources are displayed alongside one another in separate panels (cf. Figure 1). Information in all corpus panels can be retrieved through a linguistic search engine which allows, among others, the search for patterns of single words, combinations of words and combinations of words and part-of-speech patterns. It is, thus, possible to retrieve examples for multi-word-phrases (e.g. collocations) and grammatical constructions (e.g. a verb used in the passive voice).

The DeRiK corpus shall be integrated into this framework as an independent panel as well as a subcorpus of the DWDS core corpus and, thus, fill the “CMC gap” in the current version of the corpus.

The integration of a CMC reference corpus into the DWDS system will be valuable for various research and application fields, for example:

- **Lexicology and lexicography**: Besides genre-specific discourse markers and netspeak jargon (like “lol”), new vocabulary is characteristic for CMC discourse, e.g. “gru-scheln”, a form describing the virtual approaching of another person in the German so-
cial network *StudiVZ* (English paraphrase: “to poke”). The disembodiment of synchronous written communication leads to a metaphorical usage of verbs like “knuddeln” (en: “to hug sb”). These tendencies should be documented and described in up-to-date lexical resources.

- **Language variation and stylistics**: The linguistic peculiarities and the stylistic aspects of CMC are described in the CMC-related literature. However, most empirical studies on the matter have been based upon small and project-related datasets. The DeRIK corpus will provide a broader basis for qualitative and quantitative investigations on linguistic features and linguistic variation in German CMC. The DWDS framework will facilitate the comparison of CMC genres with corpora of other written genres (e.g. newspaper, fiction, scientific writing); it will, thus, be easier to investigate how new patterns and genres emerge.

- **Language teaching**: Internet communication has become an important part of everyday communication. Language- and culture-specific properties of CMC should, thus, also be regarded in communicative approaches to Second Language Teaching. In this context, the DeRIK corpus and the lexicographic documentation of CMC vocabulary in the DWDS dictionary may be useful resources. In school teaching, German native pupils may use the DWDS system to compare written language and CMC corpora and to explore how style varies across different genres (Beißwenger and Storrer 2011).

### 3 Specification of the Schema

#### 3.1 CMC Genres, Document Types, and Features Covered by the Schema

In a broader sense, computer-mediated communication comprises all communication “that takes place between human beings via the instrumentality of computers” (Herring 1996, 1). In a narrower sense, the term “computer-mediated communication” is used for such forms of communication which are based on computer networks (usually the internet). According to December (1996), those forms of computer-mediated communication can also be subsumed under the category “internet-based communication”, including all communication that “takes place on the global collection of networks that use the TCP/IP protocol suite for data exchange”. *Internet-based* computer-mediated communication can be accessed using internet or WWW client software on desktop or mobile computers as well as through applications for the use of online services on other mobile devices (mobile and smart phones).

Taking into account the focus of the DeRiK project, we restrict the focus of our schema to forms of communication which are (i) based on the TCP/IP protocol suite for data exchange, (ii) dialogic (with all participating users being able to switch between the role of a recipient/reader and the role of a producer/author of messages), and (iii) based on writing as the main encoding medium for the users’ dialogue contributions (i.e.: the verbal parts of the contributions must be encoded using writing but also may include graphics, embedded audio or video files). Thus, the present version of our schema does not cover communication which is mediated via computers while not being internet-based (such as, e.g., SMS communication) as well as monologic forms of internet-based communication (such as, e.g., monologic hypertexts) and spoken online communication such as communication on basis of audio or video conferencing software (e.g., communication via *Skype* or *Teamspeak*).

---

Our schema focuses on such forms of computer-mediated communication in which written dialogue contributions of more than one interlocutor are displayed in one and the same document. This includes communication in forums, bulletin boards, chats, instant messaging, on Twitter, on wiki talk pages, in weblogs, on user pages or discussion sections in social networks, or in online guestbooks. In those genres, the contributions of the interlocutors are first sent to a server and then put into an updated version of the document which each of the participating users can see on their computer screens and which displays the development of the ongoing conversation. The time between the receipt of a new user contribution on the server and the update of the document may vary, depending on whether in a given application contributions are inserted into the ongoing conversation immediately or whether they are first scheduled for being checked by a moderator or agent of the application provider.

In its present version, our schema excludes communication via e-mail and on the use-net in which each user contribution is stored in a separate (e-mail) document. In our opinion, the representation of documents that render only one text message (which, in addition, may have other documents in a vast range of file formats as attachments) demands a different base structure than documents which preserve sequences of contributions by two or several users. We do not exclude e-mail and usenet conversations from the DeRiK project in general; we just do not claim that the schema we describe in the following is able to cover their peculiarities adequately. Due to their differences from CMC documents that preserve chat logfiles, forum threads, Wiki discussions, and the like, e-mail and usenet documents will have to be described within an own schema which is not the subject of this paper.

The schema draft that we describe in the following sections gives a core model for the representation of the following types of CMC documents:

- threads in online forums and in bulletin boards;
- discussion threads on talk pages in Wikis;
- logfiles of conversations in webchats, on Internet Relay Chat (IRC), and in instant messaging applications;
- sequences of user postings in online guestbooks (which have a similar structure as chat or instant messaging logfiles);
- sequences of postings and threads on profile pages and in discussion sections of social network sites;
- sequences of user postings on Twitter (e.g., “timelines” of postings that include the same thematic hashtag);
- discussion threads in weblogs;
- sequences of review postings for products presented on online shopping sites;
- threads and sequences of “private messages” preserved in users’ individual mailboxes on social network sites or learning platforms;
- (etc.).

The status of our schema is that of a core model for the representation of CMC. This means that the schema is meant to provide elements for the representation of the basic structural peculiarities on the macrolevel and of some prominent linguistic features that can be found on the microlevel of CMC discourse. The structural elements on the microlevel are those elements that can be found in the content of individual users’ contributions to CMC conversations while the constituting structural elements of the macrolevel are the users’ contributions themselves. Microstructures are made of linguistic units, punctuation, media objects, and hyperlinks. The current version of our schema confines itself to selected microstructural elements which can be regarded as typical for CMC and for which annotation schemata devel-
oped for other types of discourse cannot be adopted one-to-one—especially the CMC-specific interaction signs (cf. sect. 3.5). It is self-evident that the microstructural component of the schema could be extended also on other linguistic and structural phenomena of CMC discourse (For an overview of linguistic features in German CMC discourse cf., e.g., Runkehl et al. 1998 and Storrer 2009; for English see, e.g., Crystal 2001 and the contributions in Herring 1996). The schema version which is presented in the following sections and which can be retrieved from the web at http://www.empirikom.net/bin/view/Themen/CmcTEI is open for such extensions.

3.2 Basic Modeling Decision: Customizing TEI’s Basic Formats for the Representation of Text Structure

None of the modules in the current version of the TEI Guidelines can be adopted one-to-one for creating a model for the representation of CMC. There are many elements in the default text structure module which are useful for describing the structure of individual users’ contributions to CMC discourse—but CMC documents can be regarded as text documents only in a very technical sense as they include sequences of stretches of written language which, due to their separation through line-breaks, visually appear paragraph-like. On the other hand, the dialogic structure of CMC discourse appears similar to the structure of spoken conversations—but, in contrast, the production of the users’ contributions to CMC dialogues is a monologic activity and, thus, more text-like than similar to processes of oral verbalization in which the partners perceive and process the verbal utterance simultaneously with its production. Therefore, neither the module default text structure nor the module transcribed speech nor any other module in the current TEI-P5 provides a model of interpersonal communication that fits the particularities of the main constituting elements of CMC discourse. These are the stretches of text that an individual user produces in private and then passes on to the server through performing a “posting” action (usually by hitting the [ENTER] key on the keyboard or by clicking on a [SEND] or [SUBMIT] button on the screen).

The commonalities and differences of CMC discourse with text and speech have been widely addressed in the CMC literature. CMC can best be described as (synchronous or asynchronous) written or typed conversation (Werry 1996; Storrer 2001; Beißwenger 2002) or as interactive written discourse (Ferrara et al. 1991; Werry 1996) which has to be regarded as crucially different from spoken conversation as well as from texts since it uses features of textuality for the purpose of dialogic exchange (cf. also, e.g., Crystal 2001, 25-48; Hoffmann 2004; Zitzen & Stein 2005): Just as texts, CMC is written. In some CMC genres, the users can apply text formatting features and a paragraph structuring to their contributions. In contrast to texts and similar to spoken conversation, CMC discourse is dialogic while the users’ contributions to CMC dialogues are being composed in a private activity, then sent to the server, then displayed on the screens, and it is not until then that they can be read by the other users (Beißwenger 2003, 2007). This “pre-transmission composition” protocol for the production of dialogue contributions in CMC is text-like, not speech-like. Accordingly, even in synchronous modes of CMC (chat, instant messaging) the users lack the possibility to provide simultaneous feedback or to perceive and process the contributions of their interlocutors simultaneously with their verbalization (a fact which has crucial consequences on the interactional management layer, esp. the turn taking system of conversation; cf., e.g., Garcia & Jacobs 1998, 1999; Herring 1999; Beißwenger 2003, 2007; Schönfeldt & Golato 2003; Ogura & Nishimoto 2004; Zitzen & Stein 2005). As could be shown through observations of message composition in
chats, the message production includes subprocesses of evaluation and revision (re-writing) which are particular for the production of texts (cf., e.g., the findings on message production in chats in Beißwenger 2007, 2010). All in all, CMC can, thus, be considered as more than just a hybrid of text and speech (Crystal 2001, 48). Therefore, neither text nor speech/conversation provides an adequate model for its description. Moreover, considering the form and production of user contributions to CMC conversations, a text model seems to be a better starting point for practical modeling purposes than a speech model. Or, in Crystal’s words: “On the whole, Internet language is better seen as writing which has been pulled some way in the direction of speech rather than as speech which has been written down” (Crystal 2011, 21). Still, this does not mean that written language is a good model for CMC per se; but certain structural features specific to written language can also be found in CMC, and therefore a model for the description of text can provide more elements that can be adopted for the description of written CMC than a model for speech which is bound to completely different conditions of verbalization and mutual perception.

For our schema, we decided to use the TEI header structure (P5, module 2) as the basis for the representation of metadata in CMC documents (with some minor customizations which will be described in section 3.5). For the representation of the document structure, we decided to tailor a customized version of the TEI default text structure (P5, module 4) and, additionally, of some elements from the Common Core module (module 3; esp. the p element for the annotation of paragraphs). The main issues that we had to deal with while customizing the respective TEI modules for the representation of CMC were (i) the question of how to represent the users’ written contributions as the main constituting elements of CMC conversations, (ii) the question of how to represent CMC-specific types of grouping sequences of users’ contributions to larger units (threads and logfiles), and (iii) the question of how to differentiate between the inner structure of the individual users’ contribution and the structure of the CMC discourse (the first being controlled by the user, the second being the result of an interactional achievement of all participating users or/and of a certain server routine for ordering incoming user postings).

Regarding (i), we decided to introduce a new element <posting> and assign it to the divLike class of elements (sect. 3.3.1). Regarding (ii), we decided to introduce two new <div> types and name them “thread” and “logfile” (sect. 3.3.2). Regarding (iii), we decided to use the p element for segmentations in the content of postings (CMC microstructure) and to use div elements for segmentations above the posting level (CMC macrostructures).

### 3.3 Elements of the Document Macrostructure

#### 3.3.1 The posting Element

The element posting is the basic CMC-specific element in our schema. In CMC documents it represents the largest structural units that can be assigned to one author and one point in time. The category posting is defined as a content unit that has been sent to the server “en bloc”. Its function is to make a (written) contribution to the ongoing dialogue. After being sent (“posted”) to the server, the submitted unit is displayed in the CMC document as one continuous stretch of content (text plus embedded media objects such as graphics or video files, etc.). It is usually assigned to the user name of its author (= the user who has sent the unit to the server) and often also to a certain point in time (indicated through a timestamp). Therefore, postings
can be recognized by their formal structure and, thus, be annotated automatically, even if they may have different forms and structures in different CMC genres or applications.

Fig. 2: Macrostructure of a Wikipedia talk page (excerpt).

The example given in figure 2 shows an excerpt from a Wikipedia talk page. Individual user posts all end with a signature that gives the author’s name and a timestamp. For example, the signature of posting no. 1 assigns the posting to an author named Netpilots and indicates that it was received by the server at 10:36, July 28, 2011 (CEST). More information about the author can be found on the author’s profile page, which can be accessed through the hyperlink underlying the name.

Each posting is separated from the other authors’ preceding and following posts by a leading which is larger than the standard line spacing. This makes the sequence of postings in the document appear like a sequence of paragraphs in a text documents. In addition, individual postings can contain an internal structuring. Posting 1, for example, structures its content into four paragraphs with the second and the third forming items in a bullet list. Furthermore, the author of posting 1 uses hyperlinks to connect certain segments of his posting with other Wikipedia pages (“Schwäbisch Gmünd”, “Facebook”) or with Wikipedia-external WWW resources (“Gescheiterter Bud-Spencer-Tunnel/Focus.de”, “Artikel im Tages-Anzeiger”) and bold font type to highlight the segment “Bud Spencer Tunnel” in the first paragraph.

In addition to the leading, the postings in Example 1 are also separated from each other by different levels of indentation. The indentations were deliberately added by the authors in an attempt to create thread structures, similar to those in discussion groups. Thus, the level
of indentation is a feature of the posting itself and not something that has been automatically assigned by the server.

The example given in figure 3 shows an excerpt from a chat logfile. The postings here are linearly placed one after another in the order of their arrival on the chat server. In the posting protocol on the screen, each individual post is rendered as its own division, and the server automatically adds information about the authors—the user’s nickname, which is inserted in front of every posting.

105 Dill die rosi ihr englisch ist nihct vom feinsten
rosi’s english is not the best
106 Rosenstaub1979 Nó
Nope
107 Rosenstaub1979 is schon zuuulang her
it’s been toooooo long
108 Dill aber rosi ist prächtig
but rosi is magnificent
109 Dill prachtvoll
grand
110 Rosenstaub1979 Ich glaube, so 9 Jahre
I think, about 9 years
111 Rosenstaub1979 *lol* @Dill
*lol* @Dill
112 Dill 9 jahre?
9 years?
113 Rosenstaub1979 Ja, kommt fast hin
Yes, that’s about right

Fig. 3: Sequence of postings in a chat room.

Postings represent a category in its own right which is different from elements of the text structure as well as from the constituent elements of speech: Under aspects of planning and coherence, postings exhibit similarities to spoken utterances; under the aspect of production, postings are text-like artifacts. Therefore, they may neither be identified with divisions or paragraphs in texts nor with utterances in spoken discourse. In the following, we will elaborate on this point:

According to the TEI Guidelines, the paragraph element *p* is used to mark “the fundamental organizational unit for all prose texts, being the smallest regular unit into which prose can be divided” (TEI P5: 3.1) while the element *div* identifies subdivisions of a text, e.g. chapters or sections (TEI P5: 4.1). Being defined as an “organizational unit” (of a text), the notion of the paragraph implies that there is an author or at least an author-like authority (editor, publisher) who makes certain structuring decisions while composing his text and, thus, divides it into a series of units, e.g. according to subtopics and information units. In CMC, instead, one author’s reach ends with the beginning and end of his current posting while the structure of the sequence of postings is either due to a server routine (as is the case in chat logfiles) or a joint achievement of the group of users (as is the case on Wiki talk pages or in certain forums). The resulting structure is, thus, not based on any sort of global structural planning of the text. Modeling a user posting as a paragraph would therefore reduce the original concept of the paragraph to absurdity: A paragraph is a holistic unit determined by (one author’s) global text coherence; in contrast, a posting in CMC is an atomic constituent of a written dialogue determined by the ongoing dialogue’s local coherence.
When in Example 2 chatter Rosenstaub sends posting 106 (“Nope”), she does so as a direct reaction to the previous posting 105 from user Dill. This reaction of hers was not previously determined by an author (as is the case e.g. with individual characters’ utterances in dramatic dialogues), but she reacted in this way because the previous posting created a context which made this type of response seem sensible for her locally. Before reading posting number 105, Rosenstaub could not even know herself that her own next contribution would be “Nope”; the intention for her “Nope” response is directly caused through the reception and processing of posting number 105. On the other hand, chatter Dill, when he sends his posting number 105, does not know which type of posting will follow in 106 (or if any reaction at all will come from Rosenstaub)—all because there is no author who planned the entire dialogue in advance; instead, the dialogue is developed by the users as they go along; at the same time each posting creates a context for the partners’ responses that follow. Both participants are acting according to their own communication goals; but neither of the participants can precisely predict in advance how the dialogue will really develop.

Postings also differ greatly from utterances in spoken conversation. Thus, the element \(u\) (utterance) from the TEI’s spoken module (“transcribed speech”)—describing “a stretch of speech usually preceded and followed by silence or by a change of speaker” (TEI P5: 8.3.1)—is also an inadequate option for the conceptualization of postings. The simultaneousness of verbalization, perception, and mental processing as one very central characteristic of spoken utterances is not present in postings. Due to the abovementioned “pre-transmission composition” protocol, the projection of completion points for turn-constructional units and, thus, the turn-taking apparatus do not function in the same way as in oral conversations. Postings—like texts—are first produced in entirety; the process of verbalization can accordingly not be tracked by the other participants. The new message, thus, comes to the partners as a result of the verbalization process which must be read \textit{ex retrospect}. In spoken conversation on the other hand the listener can give immediate feedback and, thus, directly react on (and affect) the ongoing verbalization; he can project transition-relevant places and negotiate turns simultaneously with the linear unfolding of the current speaker’s utterance (cf. e.g. Sacks, Schegloff and Jefferson 1974; Schegloff 2007).

In our schema, the element \textit{posting} is the basic structural element of a CMC document. It is the pivot between the higher level macrostructural elements (thread, logfile; cf. sect. 3.3.2) and the microstructure of the content which it encloses (cf. sect. 3.5). Posting belongs to the high-level elements and occupies its place in the \textit{model.divLike} class alongside the \textit{div} element.

Thus, we base the structure of the element on the structure of the existing \textit{div} element. Therefore, \textit{div} and \textit{posting} have a lot in common, but there are also quite a few differences between them which are worth mentioning.

Regarding the similarities of \textit{div} and \textit{posting}, we would like to stress the following aspects:

\begin{itemize}
  \item \textit{div} and \textit{posting} are high-level elements, belonging to the same class (\textit{model.divLike});
  \item \textit{div} and \textit{posting} contain the major divisions of text;
  \item \textit{div} and \textit{posting} have similar internal content.
\end{itemize}

It is important to note that \textit{posting}, like \textit{div}, does not belong to the class of \textit{pLike} elements. One \textit{posting} may consist of one or more paragraphs, similar to a \textit{div}. While a division may represent a chapter of a book, consisting of diverse paragraphs, \textit{posting} represents one user contribution to some computer-mediated communication event (forum, blog, web-discussion or chat). Such a contribution can contain multiple paragraphs, just like \textit{div}. In the chat exam-
ple given in figure 3, all postings consist of exactly one paragraph and the portion of text exhibits no special markups. On the Wikipedia talk page given in figure 2, some of the postings contain divisions and markup that the authors inserted into the content of their postings in order to structure their content.

Therefore, posting cannot be a model. Like element. Regarding the differences between div and posting:

- div is a self-nesting element, while posting is not;
- the occurrence of posting is restricted to a higher-level division element: postings can only appear inside of a division, which encloses one complete CMC document (e.g. an entire forum thread, or an entire blog with user comments or one chat logfile).

In other words, on one hand, posting is a child element of div, but on the other hand, it has exactly the same content model as the latter, except that it does not contain divisions and does not embed itself.

The content structure of a posting: Posting inherits the structure of the div element. Normally posting consists of one or more paragraphs. In some cases a posting contains a head, typically with a title.

The following classes of attributes can be assigned to the element posting: att.ascribed, att.datable, att.global, att.typed.

Most common attributes for posting are @synch and @who. @synch is used to signify the time when a posting arrives at the server, which processes the data and displays it on a website. These sequential points in time are ordered on a timeline. This timeline is presented separately from the postings, but in the same xml document (in the front section, cf. the code snippet in figure 4 and sect. 3.4).

The @who attribute refers to the profile of the person who submitted the posting. Profiles of all users who contributed to the conversation recorded in one CMC document are listed in the header of the xml document. The element person is used for this purpose.

Other common attributes of posting are @revisedWhen, @revisedBy, and @indentLevel. These attributes are new to TEI and were introduced to the schema through customization. The first two attributes are similar with @synch and @who but differ from them in the following aspect: they mark the time when and the person who produced a revision of a posting by editing the original content (which, in some cases, appears in Wiki and in forum discussions). These attributes take into account the fluidity of the CMC medium.

Both the @who and the @revisedBy attributes belong to the att.ascribed class. @synch and @revisedWhen are placed in the att.datable class.

The values of @synch, @who, @revisedWhen, and @revisedBy are URIs which point to a profile and to a point of a timeline respectively.

The @indentLevel attribute is newly created for our purposes and is placed in the att.global class. Its function is to mark the (relative) level of indentation of the text in a posting (as defined by its author). The values of this attribute are numbers from 1 to $\infty$ depending on the level of the indentation of the posting (cf. the encoding example given in figure 5).
Bud-Spencer-Tunnel


Original data (Wikipedia discussion)

---

Fig. 4: This example contains an encoding of a user profile, a part of the timeline, and one posting. For the complete encoding of this XML document see http://www.empirikom.net/bin/view/Themen/CmcTEI.
Freibad statt Tunnel


- Gescheiterter Bud-Spencer-Tunnel/Focus.de
- Artikel im Tages-Anzeiger Zürich


Fig. 5: Encoding of postings 1 & 2 from the example given in figure 2.

3.3.2 Threads and logfiles

As stated earlier, we use the term macrostructure to describe how series of postings are arranged in CMC documents: CMC macrostructures do not emerge from the actions of just one user but from all posting activities of all users involved in a CMC conversation plus server routines for ordering incoming user postings. The structuring on the macrostructure level of a CMC document, thus, has a different status than the structuring inserted by one and the same author into the content of his postings. In order to differentiate between divisions on the macro- and the microstructural levels of CMC, we therefore reserve the Element paragraph (p) exclusively for divisions in the content of individual postings, while we use the div-Element exclusively for the representation of divisions on the macrolevel. In addition, we differentiate between two major types of macrostructures in CMC:
1) Logfiles, which arrange the sequence of the postings in a linear chronological order based on when they reached the server (cf. the examples given in figure 7);

2) threads, which structure the sequence of postings by using two dimensions, each of them with specific semantics:
   
a) the above/below dimension, which in the standard case stands for a temporal “before/after” relation;

   b) the left/right dimension, in which one can use indentation to emphasize the topical affiliation of one message to a previous message (cf. the example given in figure 6).

For the differentiation between those two CMC-specific macrostructure types, we introduce the parameter values “thread” and “logfile” to the attribute @type of the element div.

![Fig. 6: Differentiation between CMC macro- and microstructures; CMC macrostructure type “thread”](image)
3.4 Metadata and Anonymization

3.4.1. Metadata

Metadata are used to keep record of the data to which these metadata are attached. In our context, it is convenient to add metadata to each individual document. In our case, the TEI metadata schema is sufficient to record such data which are relevant for the description of a CMC document. However, we want to draw the attention of the reader to the following features which are particular for the CMC document type:

1. On the World Wide Web, documents are quite difficult to identify. Mechanisms of persistent identifiers are currently just gaining ground and are far from being established. We therefore follow a double strategy: in cases where we are able to refer to a persistent identifier (as is the case e.g. with versions of Wikipedia talk pages), we do that as a part of the source description. In cases where we cannot refer to a persistent identifier, we download (the source of) the web page and store it as a digital image. In the source description part of the metadata we refer to this image.

2. As a part of the metadata, we store the profiles of the participants in the computer-mediated interactions included in our corpus. We construct these profiles from those data which are recoverable from the interaction. The reasons for doing so are explained below.

3. In addition, we store a timeline on which the individual users’ contributions to the documented dialogues, i.e. the postings (cf. sect. 3.3.1), are situated via the @synch attribute of the element posting. This is typically done for spoken language but is also useful for dialogic CMC. We are aware that in most cases we can only capture the point in time when a contribution is received and processed by the server, but the interesting point for purposes of documentation and analysis is the relative chronological order of contributions and not the absolute point in time.
3.4.2 Anonymization

In order to be able to distribute the collected CMC data as widely as possible, we need to anonymize the data. Our anonymization strategy shall support the following goals:

- Every user of the data shall be able to assign a certain set of postings in a CMC document (e.g. in a forum thread) to one and the same user.
- This user, however, shall not be identifiable as an individual of the “real world”.
- Despite that, some privileged, i.e. “authorized” users, shall be in a position to see and maintain the data which could be used to identify an individual person as author of postings in this thread. It might be useful to automatically or individually recover only some features of (a set of) user(s), e.g. their gender, if such data are available.

To achieve these particular goals, we perform the following steps:

- All of the recoverable personal data of a CMC participant are collected into a person profile. This profile is provided with an xml:id which is unique for the particular document. All person profiles are stored in the header of the document. Thus, they can easily be separated from the body of the document and therefore be hidden from the less privileged users of the data.
- Each posting is assigned and linked to this person profile via the xml:id, i.e. technically, an xml:id is the value of the @who attribute of the posting element.
- The xml:ids are also used to substitute instances of user names in segments of a given posting, e.g. within addressing terms (cf. sect. 3.5.1.5).

We are aware that the procedure of identifying names and maintaining person portfolios can be a time-consuming task. However, this effort is in some cases unavoidable and a necessary prerequisite for the publication and distribution of valuable data. We therefore want to ensure that a reliable anonymization strategy exists and can be used in such cases.

For an example of this strategy, please check the example in figure 4 (sect. 3.3.1).

3.5 Elements of the Document Microstructure

3.5.1 CMC-specific Types of Interaction Signs

Up to now, many assumptions about the Internet’s impact on language change have been based upon small datasets and the linguistic intuition and experience of the researchers. An annotation standard for typical elements of internet jargon—emoticons and acronyms, to name just a few—would help to investigate their usage and dissemination across (sub-)languages and digital genres on a broader empirical basis. However, there is no common terminology to classify the elements of internet jargon. Neither is there a consensus about the status of these elements in a natural language grammar framework. To fill this gap, we developed an annotation schema for these phenomena on the microstructure level of CMC documents. The basic linguistic description category of our approach is termed interaction sign; in the schema, instances of interaction signs such as emoticons, acronyms, etc. are being represented through the element interactionTerm. In the following, we will briefly introduce the category “interaction sign” and embed it into a broader grammatical framework. By means of examples, we will describe in a second step how the category and its subclasses are used for the annotation of our German reference corpus.
It is clear that the annotation schema suggested below has to be developed further and discussed within the CMC community. First and foremost, our schema serves the purpose of annotating required in the framework of the DeRiK project. Some of the subcategories may be specific for German CMC. However, the set of subcategories of interaction sign may have to be extended and adapted for other languages. In principle, we consider our proposal as a first step towards the development of an annotation standard that will facilitate cross-language, cross-genre and micro-diachronic investigations of “internet jargon” elements in CMC corpora. The schema favors a grammatical perspective, but it is open for extensions motivated through other fields of research, i.e. cultural studies or sentiment analysis.

3.5.1.1 Interaction Signs: Definition and Subclasses

Spoken discourse typically contains elements like “hm”, “well”, “oh my god”, “oops”, “wow” and the like. Grammar frameworks usually categorize them as interjections (e.g., Greenbaum 1996; McArthur et al. 1998; Blake 2008) or Interjektionen (DUDEN-4²), inserts (Biber et al. 1999; Biber et al. 2002), discourse markers (Schiffrin 1986), discourse particles or Gesprächspartikeln (DUDEN-4²). Responsive like “yes” and “no” typically occur in spoken and written dialogues.

In the system of syntactic categories of the three-volume German grammar of the Mannheim Institut für Deutsche Sprache, “Grammatik der deutschen Sprache” (Zifonun, Hoffmann & Strecker 1997, henceforth GDS), interjections and responsive are categorized as Interaktive Einheiten (henceforth IE). One important syntactic feature of IE is that they are not integrated in the sentence structure. Instead, they are often used as sentence-equivalent utterances like “nö” (“nope”) in posting 106 of the example given in figure 3 above, or they occur in front of or after the sentence boundaries (“ja, sollte eigentlich” in posting 2 of the example given in figure 2). If they occur within the sentence boundaries, they are not constituents of phrases with a syntactic function; instead, they are “thrown between” (< lat. interjectio) (see also Ehlich 1986 and Trabant 1998). In spoken discourse, IEs serve as devices for conversation management: they can be used to express reactions to a partner’s utterances or to display the speaker’s emotions. ³

Many CMC specific elements like emoticons, acronyms, and the like occur in the same positions and have similar functions like IEs in spoken discourse. It is, thus, not surprising, that grammars—if they describe them at all—classify these elements as interjections.⁹ In the STTS tagset, a standard for German part-of-speech classification¹⁰, most elements would most adequately be annotated using the POS-Tag ITJs (Interjektion) or PTKANT (Antwort-partikel); in the CLAWS2 tagset for English, they would fit into the category UH (interjection).¹¹

But this simple solution is not sufficient for corpus-based research on CMC jargon across languages, cultures, and genres. On the one hand, elements like emoticons are lan-

---

7 An online version of the GDS is available at http://hypermedia.ids-mannheim.de/; a brief description of the category interaction sign (Interaktive Einheit) can be found in module http://hypermedia.ids-mannheim.de/call/public/sysgram.ansicht?v_typ=d&v_id=370.
8 Cf. GDS (362): “Ihre Funktion besteht in der unmittelbaren (oft automatisiert ablaufenden) Lenkung von Gesprächspartnern, die sich elementar auf die laufende Handlungskooperation, Wissensverarbeitung und den Ausdruck emotionaler Befindlichkeit erstrecken kann”.
9 Cf., e.g., DUDEN-4²: §892, Ehlich (1986).
10 STTS: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html
11 CLAWS2: http://ucrel.lancs.ac.uk/claws2tags.html
guage-independent iconic signs that cannot be classified as syntactic units of natural languages in a strong narrow sense. On the other hand, iconic signs like the emoticon :-} and symbolic signs like the abbreviation *s* (= English “smile”) are often used as synonyms. All these elements share topological and functional features with natural language interjections in spoken discourse. By subsuming these “internet jargon” elements, interjections, and responsives under one category “interaction sign”, we want to account for their functional and semantic similarities (cf. figure 8).

![Diagram of interaction signs with examples](image)

**Fig. 8:** Typology of interaction signs (with examples).

In our schema, we introduce an element *interaction term* as a phrase-level element (*class model.phrase*) which encloses one or more instances of subclasses of interaction signs. The attribute class assigned to *interactionTerm* is *att.global*. In addition, we introduce elements for the following subclasses of interaction signs: first, the two subclasses of “Interaktive Einheiten” as described by the GDS (*interjection* and *responsive*) and, second, the four subclasses for elements which are typically but not exclusively used in written CMC discourse: *emoticon, interactionWord, interactionTemplate,* and *addressingTerm*. Each of the elements is assigned a set of attributes by which their occurrence in the corpus documents can be subclassified according to formal, positional, semiotic, semantic, and functional criteria. In the following, we outline the underlying basic ideas of choosing these categories and describe the properties of the elements introduced in our schema for their representation in our corpus data.
Emoticons are iconic units that are created with the keyboard. They are often used to portray facial expressions; with respect to their function they typically serve as emotion, illocution, or irony markers. Due to their iconic character, the use of emoticons is not restricted to CMC in one particular language; instead, the same emoticons can be found in CMC data in different languages. There are several systems of emoticons: besides the Western style emoticons, there are, e.g., Japanese and Korean style variants. Postings 3 and 5 in the example given in figure 2 include Japanese style emoticons (“Kawaiicons”); Western style emoticons can be found in the example given in figure 9.

Fig. 9: Postings on a Wikipedia talk page displaying instances of the (Western style) emoticons :o) and ;o) and instances of the interaction words *freu* (“happy”) and *g* (<“grin”). The combination of :o) and *freu* in posting 5 is an example of an interaction term that consists of two interaction signs.

In our schema, instances of emoticons are represented through the *emoticon* element, which is assigned to the *gLike* element class. Conventionally, elements of this class contain non-Unicode characters and glyphs. Although most emoticons are produced as a sequence of keyboard generated ASCII characters (dot, comma, colon, and the like), the resulting figure is comparable in its semiotic status to graphic characters, e.g. the so-called “smileys”. Some smiley faces are already part of Unicode, but obviously the variety of emoticons is still larger than can be captured by a set of Unicode characters. That is why we place the *emoticon* element in the class of *gLike* elements.
As to the attribute class, emoticons are provided with attributes from the `att.global` class and a number of specific attributes from other classes, such as `@style`, `@systemicFunction`, `@contextFunction`, and `@topology`. These are not yet available in the TEI standard and therefore have been introduced to it by customization.

The `@style` attribute belongs to the `att.typed` class and describes the native region of an emoticon. The value list of `@style` is currently set to `Western, Japanese, Korean, and Other`.  

`@systemicFunction` is also an `att.typed` attribute and has the following list of values: `emotionMarker:positive, emotionMarker:negative, emotionMarker:neutral, emotionMarker:unspec, responsive, ironyMarker, illocutionMarker, virtualEvent`.  

The attribute `@contextFunction` is also in the `att.typed` class and may adopt the same values as `@systemicFunction`.  

The distinction between a systemic and a context function reflects the semantic differentiation between the expression meaning and the utterance meaning of lexicalized linguistic units (cf. Löffner 2002). The idea is that, comparable to other lexemes, those types of emoticons (and also interaction words, see 3.5.2.2) which are commonly used in CMC can be assigned a general, context-independent meaning. On the web, there are quite a lot of lists displaying the “most common emoticons”, together with descriptions of their expression meaning (systemic function). Figure 10 shows an excerpt from Wikipedia’s list of Western emoticons; the left column renders types of emoticons, the right column gives short paraphrases of their (context-independent and, thus, systemic) function, as assigned by the authors.

In a given context of use, the function of an instance of a given type of emoticon may vary from its systemic function. Figure 11 shows an example (b) in which the “smiley” :-) and its variant :, which are usually assigned the systemic function of a positive emotion marker (“happy face”, see entry in figure 10), are used for marking irony. The context function of these elements in (b), thus, differs from their systemic function. In (a), instead, the context function of :) is identical with the systemic function; here, the emoticon :) is used for displaying a positive emotion (happiness about Shadok’s entering the chatroom).

The `@topology` attribute (which is a member of `att.placement`) captures the position of the emoticon relative to the text to which it belongs. Consequently, the range of values is set to `front_position`, `back_position`, `intermediate_position`, `standalone`.

<table>
<thead>
<tr>
<th>Icon</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;:D :-D :D 8-D 8D x-D xD X-D XD =-D =D =-3 =3 8-)</td>
<td>Laughing, big grin, laugh with spectacles</td>
</tr>
<tr>
<td>:-))</td>
<td>Very happy</td>
</tr>
<tr>
<td>&gt;[ ] [:-( :-( :-( &lt;&lt; &lt;&lt; &lt;- &lt;- :&lt; :&lt; &lt;&lt; &lt;&lt;</td>
<td>Frown, sad</td>
</tr>
<tr>
<td>:-[ ]</td>
<td>Angry</td>
</tr>
<tr>
<td>:-[ :]</td>
<td></td>
</tr>
<tr>
<td>:-[ ]</td>
<td>Wink, smirk</td>
</tr>
</tbody>
</table>

**Fig. 10:** List of Western emoticons as given in the English Wikipedia, page “List of emoticons” (as of 2012-02-01; excerpt).
Shadok comes in from the room Alshain.

Hey Shaddy :)

heya marc30 :o)

hey marc30 :o)

Shaddy:

Holla Shaddy :) Hey Shaddy :) heya marc30 :o)

hey marc30 :o)

Thor:

Thor... ärgerst sich immer noch, daß die franzosen den pott nicht behalten haben
Thorst is still upset that the french didn’t hold on to the pott *gg*

Erdbeere$: Erdbeere$ ärgerst sich mit .... der pott geht an frankreich und wir bekommen die küste
Erdbeere$ feels your pain .... the pott goes to france and we get the coast

Bochum:

Bochum tritt erdbeere in den arsch :))
Bochum kicks erdbeere in the butt :))

Erdbeere$:

ohh wie nett :)
ohh how nice :)

3.5.1.3 Interaction Words

Interaction words are symbolic linguistic units. Their morphologic construction is based on a word or a phrase of a given language and describes expressions, gestures, bodily actions, or virtual events—cf. the units sing, g (< grins, “grin”), fg (< fat grin), s (< smile), wildsei (“being wild”) in the example given in figure 12; they are used as emotion or illocution markers (postings 865, 876, 880), irony markers (postings 878, 879, 886) or to playfully mimic simulated bodily activity (posting 864):

WE ARE GOING TO THE EUROPEAN CUP WITHOUT GERMANY

Ryo changed colors

yep too bad

Windy123 is going to another room: Forum

all the people have to pay for their TV at media markt

Es gab mal ein Rudi Völler......es gab mal ein Rudi Völler......onkel sing

There once was a Rudi Völler......there once was a Rudi Völler......onkel sing

hehe..das wurd eh gerichtlich gestoppt juliana

hehe..that was stopped by the courts anyway juliana

echt?

really?

hehe..that was stopped by the courts anyway juliana

ja

yeah
Juliana: wieso? *why*

Gangrulez: wettbewerbsverzerrung  
*distortion of competition*

Naturkonstantler: Fussball ist sooo unendlich unwichtig...  
*Soccer is sooo incredibly unimportant...*

Juliana: versteh ich nicht. i ch fand es war ein cooler trick  
*i don’t understand. i thought it was a cool trick*

Gangrulez: aber es war eine Art Glücksspiel  
*but it was a kind of gamble*

Turnschuh: mag auch keinen Fussball......nur wollte ich das letzte Deutschlandspiel sehen  
*Turnschuh also doesn’t like soccer......but I would have liked to have seen the last Germany game*

Chris-Redfield: *s* aber net erlaubt @ juli  
*smile*

Juliana: fußball ist nen dreck wichtig. es ist ein spiel. hauptsache, die jungen männer  
*soccer isn’t worth it. it’s a game. Main thing, the young men have kept fit and  
done something for their health :)*

Gangrulez: und das entspircht nicht dem Handel  
*and that wasn’t the deal* 

Juliana: chris, du weißt doch, daß ich ein gesetzesbrecher bin  
*smile*

Chris-Redfield: ja ich weiß *s*  
*yes i know* 

Juliana: *being wild*  
*naja... äh.
*oh well... um.*

Gangrulez: ach ich muss ja noch ne mail schreiben.  
*oh i have to write an e-mail.* 

Juliana: ich geh zu meinem buch und...  
*i’m going to go to my book and...*

System: Gangrulez geht in einen anderen Raum: sphere  
*Gangrulez goes to another room: sphere*

Naturkonstantler: vielleicht können wir ja mal eine Greencard für potentielle Fussballspieler einführen... ich werde eine Petition bein B-tag einreichen... Ja, so bin ich, ich sorge mich um das Wohl der Allgemeinheit!  
*might we can introduce a green card one day for potential soccer players... I  
will submit a petition to congress... Yes, that’s how I am, I care for society’s  
well-being!* 

Juliana: mal schaun  
*we’ll see*

System: juliana verlässt den Raum  
*juliana leaves the room*

---

Fig. 12: Excerpt of a social chat displaying instances of interaction words (postings 864, 865, 875, 876, 878, 879, 880, 881, 886) and of addressing terms (868, 876).

The element interactionWord in our schema is a member of model.global.spoken. In some way interactionWord is similar to kinesic, incident, and vocal elements. The element interactionWord is provided with attributes from the class att.global and several additional attributes. Attributes specific for this element are @formType, @systemicFunction, @contextFunction, @topology, and @semioticSource. All of the above listed attributes are new attributes of our
customized schema. The values of @systemicFunction, @contextFunction, @topology have already been described in sect. 3.5.1.2 about the emoticon element. @formType is in the att.typed class of attributes and is used to describe morphological properties of the interactionWord. Therefore, the list of values is currently set to: simple, complex, and abbreviated. @semioticSource is in the att.typed class of attributes and is used to describe the semiotic mode that forms the basis for an interaction word; its current list of values is set to mimic (as, e.g., in grins “grin”, stirnrunzel “frown”), gesture (as in kopfschüttel “shake head”, wink “wave”), bodilyReaction (as in schluck “gulp”, seufz “sigh”, hüstel “little cough”), sound (as in plätscher “splash”, blubb ”plop”), action (including linguistic actions; see tanz “dancing”, knuddle “cuddling”, erklä“explaining”, mampf“munching”), sentiment (as in freu “happy”), process (as in träum “dreaming”), and emotion (as in schäm “ashamed”).

536 Thor: Thor... ärgert sich immer noch, daß die franzosen den pott nicht behalten haben *gg*
544 Erdbeere$: Erdbeere$ ärgert sich mit ..... der pott geht an frankreich und wir bekommen die küste
554 Bochum: Bochum tritt erdbeere in den arsch :))
564 Erdbeere$: ohh wie nett :)

<p>Thor... ärgert sich immer noch, daß die franzosen den pott nicht behalten haben</p>

<interactionTerm>
  <interactionWord formType="abbreviated" systemicFunction="ironyMarker" contextFunction="ironyMarker" semioticSource="mimic" topology="back_position">
    *gg*
  </interactionWord>
</interactionTerm>

<br/>

<p>Erdbeere$ ärgert sich mit ..... der pott geht an frankreich und wir bekommen die küste</p>

<posting synch="#t544" who="#A02">
  <p>Erdbeere$ ärgert sich mit ..... der pott geht an frankreich und wir bekommen die küste</p>
</posting>

<p>Bochum tritt erdbeere in den arsch</p>

<emoticon style="Western" systemicFunction="emotionMarker:positive" contextFunction="ironyMarker" topology="back_position">:-))</emoticon>

<br/>

<p>ohh wie nett</p>

<posting synch="#t554" who="#A03">
  <interjection>ohh</interjection>
  <p>wie nett</p>
  <emoticon style="Western" systemicFunction="emotionMarker:positive" contextFunction="ironyMarker" topology="back_position">:-))</emoticon>
</posting>

<br/>

<p>Erdbeere$:</p>

<posting synch="#t564" who="#A02">
  <p>ohh wie nett</p>
  <emoticon style="Western" systemicFunction="emotionMarker:positive" contextFunction="ironyMarker" topology="back_position">:-))</emoticon>
</posting>

<br/>

Fig. 13: Encoding snippet for example 11b from figure 11.
3.5.1.4 Interaction Templates

Interaction templates are units that the user does not generate with the keyboard but by activating a template, which then automatically inserts a previously prepared text or graphical element into a space of the user’s choice.

Amongst others, the category of interaction templates includes graphic smileys which the user of a CMC environment can choose from a finite list of elements. These often portray not just facial expressions but can depict almost anything; in the case of animated *gif graphics, they can even portray entire scenes as moving pictures. This clearly goes beyond that which can be expressed using only keyboard-generated emoticons. On the other hand, keyboard-generated units can individually be varied, and users can even invent new forms, while template-generated units are always bound to predefined templates.

The element interactionTemplate in our schema belongs to the model.global class of elements. It is provided with the att.global class of attributes and a few new attributes which belong to different classes. The most important attributes for this element are @type, @motion, @systemicFunction, and @contextFunction.

As the attribute @type is used to characterize the surface of the figure, the list of values is currently set to: iconic, verbal, and iconic-verbal.

The @motion attribute belongs to the att.typed class and describes yet another surface feature of interaction templates, namely whether it is a static or an animated image. Therefore, there are two types of values for this attribute: static and animated.

The attributes @systemicFunction and @contextFunction have already been introduced in sect. 3.5.1.2. Therefore, only one additional value of attribute @systemicFunction should be mentioned. The value “evaluation” is used to express whether the enclosed graphic element expresses appreciation or disapproval.

3.5.1.5 Addressing Terms

Addressing terms are units which are used to address an utterance to a particular interlocutor (see the examples in the postings 868 and 876 in figure 12). The most widely used form here is the one made out of the <@> character together with a specification of the addressee’s name.

The element addressingTerm in our schema belongs to the model.nameLike class of elements. This element is usually not specified by any attributes; nevertheless, the provided attribute class for it is att.global. The content of addressingTerm is restricted to two elements: addressMarker and addressee.

The addressMarker element belongs to the class model.labelLike and is provided with the att.global class of attributes. LabelLike elements are used to gloss or explain parts of a document. In particular, the purpose of addressMarker is to identify or to highlight the addressee in a posting. This is typically achieved by using the ‘at’-sign (‘@’) or one of a set of fixed phrases (E: ‘to’; G: ‘an’, ‘für’).

The element addressee is placed in the model.nameLike.agent class. The attributes assigned to this element are @who, @scope, @formType, and attributes from the att.global class.

The addressees are often addressed using abbreviated or nickname forms of their user names. The name of the addressee given in the addressing term is then not identical with the user name of the respective interlocutor. We would like to enable the users of our corpus to
retrieve this information from the data even after the corpus data have been anonymized (cf. sect. 3.4). We use the \texttt{@formType} attribute for this purpose and assign it the following set of values: \texttt{persNameFull}, \texttt{persNameAbbreviation}, \texttt{persNameNickname}. Thus, the attribute \texttt{@formType} allows us to describe cases like the ones illustrated through the examples in figure 14:

14a:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
|306| **Lantonie** | Lantonie heiratet Thor....  
\textit{Lantonie is marrying Thor....} |
|308| **Lantonie** | :))  
:)) |
|323| **zora** | wos? *eifersüchtel*\texttt{@lanto}  
\texttt{what? *jealous*}@lanto |

14b:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
|104| **Chris-Redfield** | tom ram ist doch nicht alles im leben \*g*  
tom ram is not all there is in life \*g* |
|108| **TomcatMJ** | nö, aber hilft dem server weiter@c-r  
\texttt{-)}  
\texttt{no, but helps the server@c-r  
\*)} |

14c:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
|117| **Raebchen** | Raebchen rät allen Pärchen, nicht auf Deck zu knutschen (sowas hat die Titanic sinken lassen! habe ich im Film gesehen)  
\textit{Raebchen advises all couples not to make out on deck (that’s what made the Titanic sink! i saw it in the movie)} |
|123| **McMike** | *lol*\texttt{@Raeby}  
*lol*\texttt{@Raeby} |

14d:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
|89| **McMike** | könntet Ihr mich bitte zum Käpten ernennen?  
\textit{could you all please appoint me captain?} |
|94| **ineli26** | ineli26 ernennt McMike zum Kapitaen  
Ineli26 appoints McMike captain |
|   |   |   |
|160| **McMike** | Monk, kannst Du das steuer übernehmen?  
\textit{Monk, can you take over the wheel?} |
|164| **Monk** | klar wohin solls gehen?  
of course where to? |
|169| **McMike** | Monk immer dem Fön nach  
\textit{Monk keep following the Foen} |
|172| **ineli26** | lol @ \texttt{kapitaen}  
\texttt{lol @ kapitaen} |

\textbf{Fig. 14:} Types of variation of addressees’ names in addressing terms: 14a and 14b: abbreviated form, 14c and 14d: nickname form (excerpts from documents no. 2221006, 2221007, and 2221001 in the Dortmund Chat Corpus).

The \texttt{@scope} attribute belongs to the \texttt{att.scoping} class. This attribute is used to specify whether one or more persons or groups are addressed. For this reason the values of this attribute are: \texttt{all}, \texttt{group}, \texttt{individual}, \texttt{unspec}.

The \texttt{@who} attribute is supposed to mark the name of the addressee, i.e. of the recipient. As to the value of \texttt{@who}, it always points at the xml:id of the person to whom the message is addressed\textsuperscript{12}.

Figure 15 gives an encoding example for addressing terms in chat postings.

\textsuperscript{12} This is part of the anonymization strategy, cf. section 3.4 for details.
3.5.2 User Signatures

An important element of the microstructure in postings in forums, bulletin boards, and Wiki discussions is the signature text which is predefined by the user and inserted into a posting automatically (usually at its end). It often includes the name of the user plus additional text (e.g., sayings, proverbs, quotes, personal information about the user) and graphics. In our schema, we do not represent signatures as a part of every single posting; instead, we mark the position in the posting where the user signature is placed and describe its content only once in the user profile in the document header.

For the representation of the signature text’s position in the postings and for the description of the signature content, we introduce two special elements: The element auto-Signature is an empty element which is contained in the model.pPart.edit class. It replaces the signature text in the posting. The user’s signature is kept in the element signatureContent in the header of the user profile; it is placed in the model.persStateLike class and pointed at within autoSignature by using the @target attribute.

3.5.3 Postscripts, Openers, and Closers

Some elements in CMC discourse are similar to elements known from written letters. Their use is, however, less restricted than it is with their functional equivalents in written letters.

Fig. 15: Encoding snippet for the postings 868 and 876 from the example in figure 12.
One element of this type is the postscript. In CMC, a complete posting can be marked as a postscript (e.g., by introducing it with “p.s.”); in other cases, a postscript can be a part of a paragraph (cf. the examples given in figure 16). The current TEI definition of the postscript element does not offer any opportunity to encode such cases. In our schema, we therefore introduced a <seg type="postscript"> for their annotation.

16a:

p.s.: ich hasse einfache antworten deshalb würde ich die antwort von <<user2>> kritisieren wollen: warum ist der "normal-christliche" lebensstil in so feste bahnen zementiert? warum läuft es trotzdem so schief. […]

p.s.: i hate simple answers which is why I would like to criticize the answer given by <<user2>>: why is the "normal Christian" lifestyle so strictly regulated? Why despite this does is still go wrong. […]

(Follow-up message of user1 to his own prior posting in a blog discussion; anonymized)

16b:


The mentioned sources are in no way trustworthy for this question, i.e. it would be conspiracy theory. In Volkach the Main Bridge is only called the Main Bridge because there is only the one for the locals. But the owner, the state of Bavaria, of course, has several Main bridges, making this one the Main Bridge Volkach. Thus, this construction will definitely not be called Bike Bridge, you would have to ask at the City of Constance’s planning department. Otherwise, stick with the same terminology as in the more respectable literature, Geh- und Radwegbrücke über den Seerhein bei Konstanz. --Störfix 21:55, 13. Jul. 2011 (CEST) P.S. or finally name the bridge after a deserving mayor ;-)  

(Wikipedia talk page for the article “Geh- und Radwegbrücke über den Seerhein bei Konstanz”)

Fig. 16: Types of postscripts in CMC: 16a: postscript posting, 16b: postscript as part of a paragraph (within a posting).

CMC communication is characterized by a less conventional style of writing. This affects also the form of a posting. We assume that, similar to conventional discourse types such as letters, some kinds of postings (especially in asynchronous CMC genres such as forums, bulletin boards, and Wiki discussions) have a structure which consists of an opening part, the main part of a message, and a closing part. However, the opening and closing parts are in many cases neither cleanly separated from the body of the message nor are they the first / last part of the message (see example below). Additionally, an opener or closer element can appear more than once in a posting.

Unfortunately, the elements of the current TEI framework (P5) which come closest to these structures, i.e. the opener and closer elements, have a too restricted distribution. For example, the element opener may appear exclusively at the top of a division, while closer is permitted at the bottom of a document only. We are ready to make use of the established opener and closer elements also for CMC documents, but the TEI rules and restrictions which hold for these elements would have to be changed first to allow for a more liberal distribution of these elements. For example, it would be useful if the opener and closer elements could join the inter-level elements. Thus, they would be able to appear within as well as in between the chunks of text. In the current version of our schema, we use seg elements for the annota-
tion of openers and closers in CMC postings and type them as “opener” and “closer” respectively (see the example given in figure 17).

Fig. 17: Opener and closer inside one posting, encoded using the seg element.

4 Conclusions and Outlook

We have shown in this paper that the TEI Guidelines offer an appropriate way of structurally encoding documents of various CMC genres. We exemplified this by focusing on some of these genres – chats, forum, and Wiki discussions, in particular – and on some features of dialogic CMC which have figured prominently in the linguistic literature about this text type.

Customization of the TEI Guidelines is one way of adapting the TEI encoding framework to new genres and document types. However, regarding the relevance of CMC in today’s everyday communication, it could be an important extension to future versions of the TEI Guidelines to include a standard for the representation of the features and peculiarities of CMC genres and document types. On the one hand, such a standard should include a model for the representation of those structural and linguistic features of CMC discourse which are not yet covered by the modules and elements in the P5 version of the TEI Guidelines (amongst others: a posting element for representing the main constituting units of the CMC document structure and elements for the annotation of typical “internet jargon” units such as the interaction signs described in sect. 3.5.1). On the other hand, a standard for the representation of CMC discourse should take into account that the distribution and content model of certain elements from existing modules in TEI-P5 would have to be modified in order to use them for the annotation of their functional equivalents in CMC postings. As shown in the example of postscript-, opener-, and closer-like elements in CMC, (cf. sect. 3.5.2), the role model of written letters that some CMC users adopt for the design of their postings is somewhat loosened in CMC; accordingly, the position of these elements in the structure of the postings is less restricted than in “traditional” written letters. In cases like these, a modification of existing TEI elements (here: the elements postscript, opener, and closer from the TEI-P5 text module) would ideally best account for both aspects of CMC phenomena: their orientation toward traditional text types and text elements as well as their free and creative use and modification.

CMC is constantly gaining popularity, both as a medium of communication and as an object of study. We therefore want to suggest with this paper that the TEI offers users a framework for annotating resources of this type. We hope that the schema presented here could pave the ground for such a development.

Much still has to be done to achieve a fuller understanding of CMC genres and their peculiarities. This is not due to a lack of studies of this kind of communication, but to a con-
stant change both in the ways in which the medium is used and in its technological frameworks. CMC is a fluid mode of communication, and we probably will have to constantly adapt our modeling and schema to new forms and ways of CMC which will emerge in the future. We are confident that the TEI guidelines will provide an appropriate framework for this. CMC is also a multilingual mode of communication (Danet and Herring 2007). We hope that the further discussion of the schema presented in this paper will help uncover the extent to which its core features can be appropriate for the representation of CMC discourse in languages other than German (and especially those with writing systems different from the Latin alphabet).

According to the DWDS project context (building a balanced general language corpus of contemporary German), all data collected for the DeRiK corpus will be German. For DeRiK, we are facing the following challenges in the near future:

1. **Acquiring texts in larger proportions:** Up to now we have been working with a small sample of texts of various genres. In the future we will acquire a larger set of documents for our reference corpus, ideally a text volume of 10 million tokens per year. We have to clear the rights of many of the text sources, if they have not already been cleared by the providers, which is the case e.g. with Wikipedia talk pages. We hope that we can acquire substantial portions of data from projects focused on empirical research in the field of CMC (amongst others, the projects from partners in the “Empirikom” network). Ideally, this would be a win-win situation. The partners would get their texts curated and distributed in a way that the empirical basis of their research could be used to replicate their work or to perform comparable research on the same data. On the other hand, more users and researchers could find and use these data easily.

2. **Analyzing CMC texts linguistically:** The software for the automatic analysis and annotation of texts is optimized for well-formed written clauses and sentences. CMC texts will therefore pose challenges to these tools on different levels, from tokenization (‘sachma’ (one string) = ‘sag mal’ (two tokens)) and sentence boundary detection to part-of-speech tagging and syntactic parsing. We hope to have shown with the examples in this text that, seen from the perspective of a normative grammar for written text, many productions of CMC are not ‘well-formed’. It will be a major challenge to find and describe the regularities in text production which seem to be irregular at first sight. NLP tools have to be adjusted accordingly. Of course there is a continuum ranging from well-thought out – and formulated – texts and dialogues (e.g. on Wikipedia talk pages or scientific blogs) to very informal and highly speech-like contributions in some chats. Tools for the linguistic analysis of CMC should be able to cover the whole range.

3. **Annotating the collected data using our TEI schema:** Last but not least, the data collected for integration in our corpus will be annotated using the schema presented in this paper. We assume that some of its structure can be generated automatically on the basis of filters that transform structural patterns of the raw data format (e.g., HTML) into the target format; other components of the schema (especially the functional subclassification of types of interaction signs using attributes), will, at least in the beginning, require manual or at best semi-automatic encoding. Further analyses of CMC-specific units on the microlevel of postings may help to develop strategies for a partial automatization of this task; we hope that the further discussions in the context of the Empirikom network will contribute to this.
Providing a framework for managing a corpus of CMC data: Scripts will be needed to i) transform CMC data of various sources to the TEI target format; ideally this will be a framework which can be parametrized for each individual source; ii) transform the TEI/XML-encoded data into something which can be displayed nicely; XSLT-scripts will be an appropriate means. We will provide such scripts and tools alongside the schema and documentation on our website (cf. FN 1). Additional facilities will be provided by the DWDS framework (cf. Section 2.2).

5 References

5.1 Works cited


Konrad Ehlich, Interjektionen (Tübingen: Niemeyer, 1986).


Gisela Zifonun, Ludger Hoffmann und Bruno Strecker, Grammatik der deutschen Sprache, 3 vols., Schriften des Instituts für deutsche Sprache 7.1-7.3 (Berlin: de Gruyter, 1997)

5.2 WWW resources

Author info

**Michael Beißwenger** is a researcher and lecturer for German linguistics at TU Dortmund University and coordinator of the DFG scientific network “Empirical research on Internet-based Communication” (http://www.empirikom.net).

Dr. Michael Beißwenger  
Department of German Language and Literature  
TU Dortmund University  
D-44221 Dortmund.  
E-Mail: michael.beisswenger@uni-dortmund.de  
Web: http://www.michael-beisswenger.de

**Maria Ermakova** is studying Historical Linguistics (M.A.) at Humboldt University Berlin. Since 2010 she is working as research assistant in the project „Digital Dictionary of the German language“ (DWDS) at the Berlin-Brandenburg Academy of Sciences (BBAW).

Maria Ermakova  
Berlin-Brandenburgische Akademie der Wissenschaften  
Jägerstr. 22/23  
D-10117 Berlin  
E-Mail: ermakovamd@googlemail.com

**Alexander Geyken** is a researcher at the Berlin-Brandenburg Academy of Sciences (BBAW) where he is head of the project group of the „Digital Dictionary of German language“ (DWDS), a long-term project of the BBAW.

Dr. Alexander Geyken  
Berlin-Brandenburgische Akademie der Wissenschaften  
Jägerstr. 22/23  
D-10117 Berlin  
E-Mail: geyken@bbaw.de

**Lothar Lemnitzer** is a lexicographer and corpus linguist at the “Digitales Wörterbuch der deutschen Sprache” (DWDS), Berlin-Brandenburg Academy of Sciences and the Humanities.

Dr. Lothar Lemnitzer  
Berlin-Brandenburgische Akademie der Wissenschaften  
Jägerstr. 22/23  
D-10117 Berlin  
E-Mail: lemnitzer@bbaw.de

**Angelika Storrer** is professor for German linguistics at TU Dortmund University. As a member of the Berlin-Brandenburg Academy of Sciences (BBAW) she is involved in the work on the DWDS project (www.dwds.de).

Prof. Dr. Angelika Storrer  
Department of German Language and Literature  
TU Dortmund University  
D-44221 Dortmund  
E-Mail: angelika.storrer@uni-dortmund.de  
Web: http://www.angelika-storrer.de