

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer and Angelika Storrer

A TEI Schema for the Representation of Computer-mediated Communication

Warning

The contents of this site is subject to the French law on intellectual property and is the exclusive property of the publisher.

The works on this site can be accessed and reproduced on paper or digital media, provided that they are strictly used for personal, scientific or educational purposes excluding any commercial exploitation. Reproduction must necessarily mention the editor, the journal name, the author and the document reference.

Any other reproduction is strictly forbidden without permission of the publisher, except in cases provided by legislation in force in France.

revues.org

Revues.org is a platform for journals in the humanites and social sciences run by the CLEO, Centre for open electronic publishing (CNRS, EHESS, UP, UAPV).

Electronic reference

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer and Angelika Storrer, « A TEI Schema for the Representation of Computer-mediated Communication », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 05 November 2012. URL : <http://jtei.revues.org/476> ; DOI : 10.4000/jtei.476

Publisher: Text Encoding Initiative Consortium
<http://jtei.revues.org>
<http://www.revues.org>

Document available online on:

<http://jtei.revues.org/476>

Document automatically generated on 05 November 2012.

TEI Consortium 2012 (Creative Commons Attribution-NoDerivs 3.0 Unported License)

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer and Angelika Storrer

A TEI Schema for the Representation of Computer-mediated Communication

1. Introduction

- 1 In the past three decades, computer networks and especially the Internet have brought forth new and emerging genres of interpersonal communication which are the subject of research in the field of “computer-mediated communication” (henceforth *CMC*). In general, genres such as e-mail, online forums, chats, instant messaging, or weblogs stand in the tradition of well-known genres such as spoken conversations or written letters. On the other hand, they display linguistic and structural features which differ from both speech and written text (see below for details) and which can be traced back to the ways in which interlocutors adapt to the technical potentials and limitations of computer-mediated communication.
- 2 Recent surveys on the use of the Internet (such as “ARD/ZDF-Onlinestudie”,¹ conducted annually in Germany) show that use of CMC applications is an important part of everyday communication. To gain 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 no longer a completely new field of research within the humanities, such a format still does not exist.
- 3 In this paper, we present an XML schema for the representation of genres of computer-mediated communication that is conformant with the encoding framework defined by the TEI. Up to now, the encoding of CMC genres and document types has not been a focus of the TEI. 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 to extend support to these genres and document types. The focus of the schema is on those CMC genres which are *written* and *dialogic*—threads in forums and bulletin boards, chat and instant messaging conversations, wiki talk pages, weblog discussions, microblogging on Twitter, and conversations on “social network” sites. The schema has been developed in the context of the project “Deutsches Referenzkorpus zur internetbasierten Kommunikation” (DeRiK, Beißwenger et al. 2012),² 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 in 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)³ at the BBAW “Zentrum Sprache”.
- 4 Since all corpus resources of the DWDS project are already encoded according to the TEI encoding framework, and since there is not yet a common standard for an XML/TEI representation of the structural and linguistic properties of CMC resources, the project group decided that the TEI would be an optimal basis for the annotation of the DeRiK data—assuming that the encoding framework of the TEI would prove 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; especially that the interlocutors’ contributions to conversations in forums,

chats, 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 to anyone 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* (including emoticons, interaction words, and addressing terms).
- It should be open to 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 toward surface phenomena and thus be as independent as possible from any specific theory of CMC discourse; this will allow use of the macrostructure model of the schema as a basic document structure in as many projects as possible; in addition, it will allow automation 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 should 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 should 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).

5 First we will give an outline of the motivation and context of the project. 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.⁴ The schema itself, its documentation, and some encoded example documents can be found online.⁵

6 The current version of the schema will form the foundation of 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 future version of the TEI Guidelines.

2. Motivation and Project Background

2.1. Motivation

7 The motivation for building a corpus of German CMC is to close a gap in the range of corpora currently available for the study of CMC and contemporary German in general. Hardly any annotated specialized corpora of CMC exist, and general corpora of contemporary German do not systematically include language as used on the Internet (Beißwenger and Storrer 2008). This poses a blatant gap since online communication has become an important part of everyday communication and can no longer be ignored when documenting contemporary everyday language use. The field of corpus linguistics is aware of that gap. In addition to the DeRiK project, which aims to build a German CMC corpus and integrate 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 including a subcorpus of CMC (Reynaert et al. 2010).

8 Due to a lack of standards for representing CMC, up to now corpus-based research projects focusing on features of CMC discourse have typically developed their own, project-specific encoding schemas (see, for example, the XML encoding for chats that has been designed for the resources included in the *Dortmund Chat Corpus*, 2003–2009).⁶ This complicates, maybe even makes impossible, the sharing of this data across projects, which 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 person hours required to annotate the data). The potential for sharing, merging, and comparing corpora, particularly in contrastive linguistic research, calls 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 already widely used in existing text and speech corpora. This would allow the schema to not only meet the needs of scholars interested in CMC but also those interested in phenomena of contemporary language in general or in comparative analyses of linguistic phenomena in CMC corpora or 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 this. As will be shown in section 3 of this paper, TEI has the power and flexibility to describe CMC structures and features even though modules and elements covering the particularities of CMC discourse are not yet implemented in the TEI. Therefore, a TEI-compliant XML schema for CMC discourse requires additional modules. Considering the relevance of the Internet as a communication medium, a separate module for CMC document types and features could be an important extension for a future version 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 should be sure to include all relevant types of genres in which the contemporary use of this language is embodied. Nowadays, for a language like German with a strong online presence, this should include genres of computer-mediated communication. In the project *Deutsches Referenzkorpus zur internetbasierten Kommunikation* (DeRiK),⁷ we are aiming to build a corpus of German CMC covering data from the most popular CMC genres. Data sampling is guided by the findings of the *ARD/ZDF-Onlinestudie*, which shows the popularity of various genres among German online users. For practical reasons, though, the project will sample only those domains and genres that are cleared from intellectual property rights. The data will be integrated in and presented through the DWDS, 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):

1. Lexical resources: a common language dictionary,⁸ an etymological dictionary, and a thesaurus;
2. Corpus resources: a balanced reference corpus (called the “DWDS core corpus”) of German from 1900 to the present. The corpus is balanced among nearly equal shares of journalistic texts, scientific prose, functional texts, and fiction. Until recently, CMC did not play a role either as an independent text genre or as part of one or more of these genres; additionally, a set of newspaper corpora and specialized corpora that are not part of the DWDS core corpus (such as German newspapers from Jewish communities edited in the first decades of the 20th century);
3. Statistical resources for words and word combinations.

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

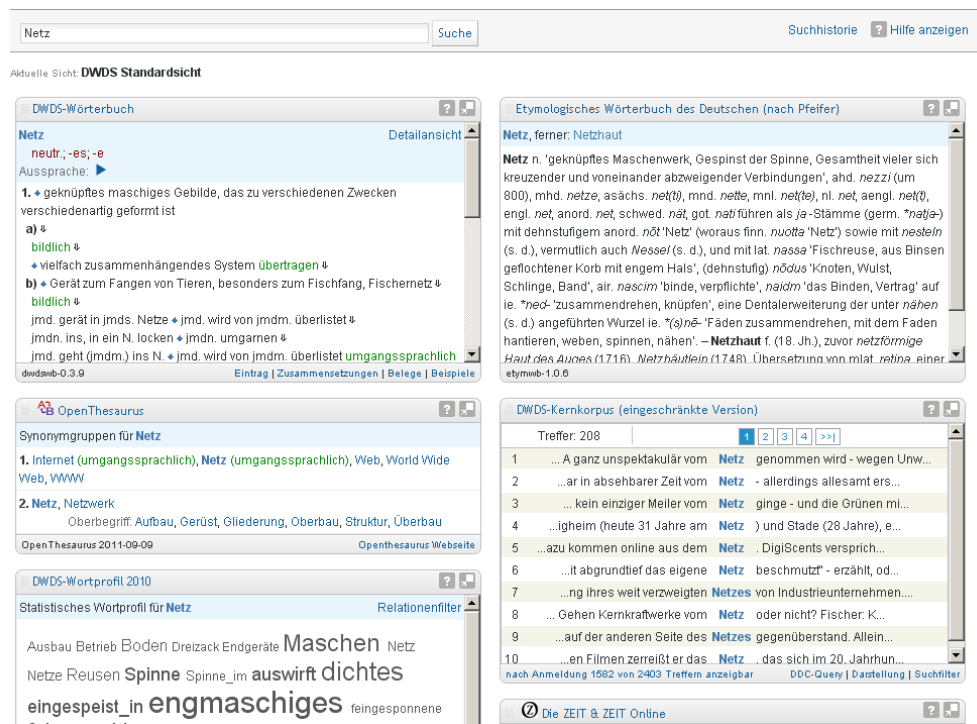


Figure 1: Web interface of the DWDS system

13 The DeRiK corpus will 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.

14 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 Internet jargon (like “lol”), new vocabulary is characteristic of CMC discourse. For example, “gruscheln”, a form describing the virtual approaching of another person in the German social network *StudiVZ* (English paraphrase: “to poke”). Furthermore, the disembodiment of synchronous written communication leads to a metaphorical usage of verbs like “knuddeln” (en: “to hug [somebody]”). These features should be documented and described in lexical resources.
- *Language variation and stylistics*: The linguistic peculiarities and the stylistic aspects of CMC are described in the CMC-related literature.⁹ 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; 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. Thus, language- and culture-specific properties of CMC should 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

15 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 that are based on computer *networks* (usually the Internet). According to John 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 communication can be accessed using client software on desktop or mobile computers or through applications for the use of online services on mobile communication devices such as mobile and smart phones.

16 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 (that is, the verbal parts of the contributions must be encoded using writing, though they may also 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 SMS communication), monologic forms of Internet-based communication (such as static webpages), or spoken online communication using audio or video conferencing software (such as *Skype* or *Teamspeak*).

17 Our schema focuses on those forms of computer-mediated communication in which written dialogue contributions of more than one interlocutor are displayed in the same document. In its present version, the schema excludes communication via e-mail and on Usenet 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 more users. We do not exclude e-mail and Usenet conversations from the DeRiK project in general; we simply do not claim that the schema we describe below is able to adequately cover their features.

18 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 structure similar to 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 (such as “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.

19 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. Structures on the microlevel (or *microstructures*) are made of linguistic units, punctuation, media objects, and hyperlinks. The current version of our schema confines itself to those microstructural elements that can be regarded as typical for CMC—especially the CMC-specific *interaction signs* (section 3.5 below). The schema could be extended in such a way that it covers further linguistic and structural phenomena of CMC discourse (for an overview

of linguistic features in German CMC discourse, see, for example, Runkehl et al. [1998] and Storrer [2009]; for English, see, for example, Crystal [2001] and the contributions in Herring [1996]). The schema presented in the following sections is open to such extensions.

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

- 20 None of the modules in the current version of the TEI Guidelines can be adopted “as is” 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 since they include stretches of written language which, due to their separation through line-breaks, appear paragraph-like. On the other hand, the dialogic structure of CMC discourse appears similar to the structure of spoken conversations (covered by the *transcribed speech* module), but the production of the users’ contributions to CMC dialogues is a monologic activity and, thus, more *text-like* than speech, in which the interlocutor perceives and processes the verbal utterance nearly simultaneously with its production by the speaker. Therefore, neither of these modules, nor any other module in 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).
- 21 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 (see also, for example, Crystal 2001, 25–48; Hoffmann 2004; Zitzen and Stein 2005): Just like text, CMC is written. In some CMC genres, the users can apply text formatting features and 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; it is not until then that they can be read by 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 and 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 (which has crucial consequences for the interactional management layer, especially turn-taking in conversation; see, for example, Garcia and Jacobs 1998, 1999; Herring 1999; Beißwenger 2003, 2007; Schönfeldt and Golato 2003; Ogura and Nishimoto 2004; Zitzen and Stein 2005). As can be seen by observing message composition in chat sessions, the message production includes subprocesses of evaluation and revision (re-writing) which are particular to the production of text (see, for example, 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 provides an adequate model for its description. But 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, “[o]n 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” (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.

- 22 For our schema, we decided to use the TEI header module in P5 as the basis for the representation of metadata in CMC documents (with some minor customizations which will be described in section 3.5 below). For the representation of the document structure, we decided to tailor a customized version of the TEI *default text structure* module and, additionally, of some elements from the *common core* module (especially 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 *logfile*s), 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 and/or of a certain server routine for ordering incoming user postings).
- 23 Regarding (i), we decided to introduce a new element <posting> and assign it to the divLike class of elements (section 3.3.1 below). Regarding (ii), we decided to introduce two new <div> types and name them *thread* and *logfile* (section 3.3.2 below). 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

- 24 The element <posting> is the basic CMC-specific element in our schema. In CMC documents it represents the largest structural unit 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.

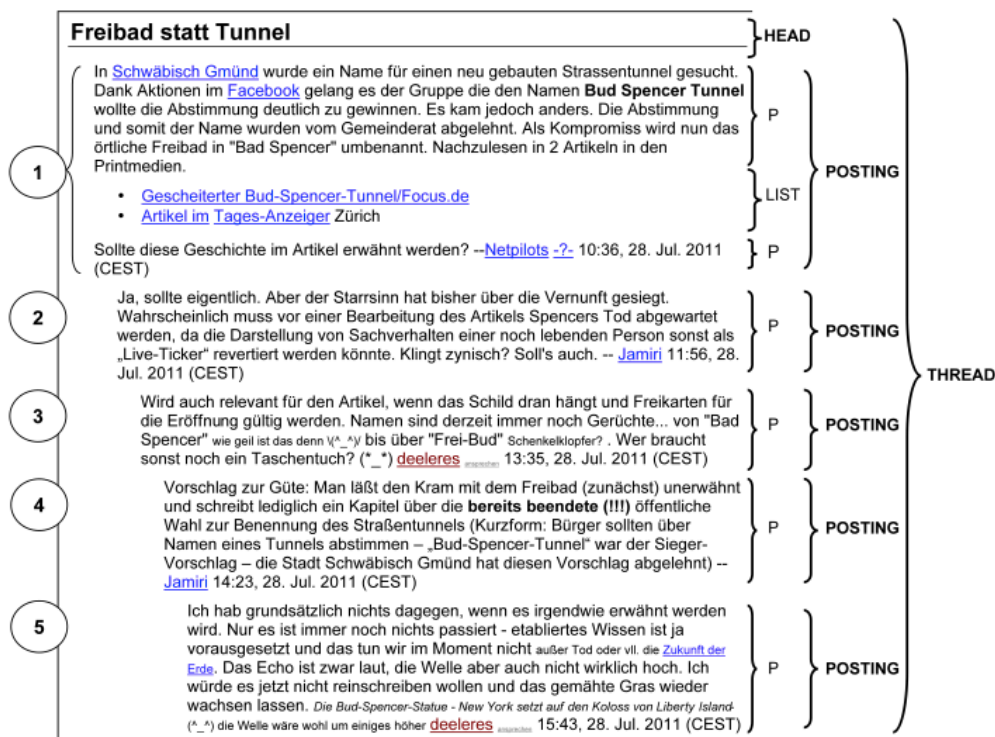


Figure 2: Macrostructure of a Wikipedia talk page (excerpt)

- 25 The example given in figure 2 shows an excerpt from a Wikipedia talk page. Individual user postings all end with a signature that gives the author's name and a timestamp. For example, the signature of posting 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.
- 26 In a Wikipedia talk page, there is a convention to use a paragraph break to separate each author's posting. This makes the sequence of postings in the document appear like a sequence of paragraphs in a text document. In addition, individual postings can have internal structure. Posting 1, for example, structures its content into two paragraphs and a bullet list with two items. Furthermore, the author of posting 1 uses hyperlinks to connect certain segments of his posting with other Wikipedia pages ("Schwäbisch Gmünd" and "Facebook") and with Web resources external to Wikipedia ("Gescheiterter Bud-Spencer-Tunnel/Focus.de" and "Artikel im Tages-Anzeiger"), plus bold font weight to highlight the segment "Bud Spencer Tunnel" in the first paragraph.
- 27 In addition to the paragraph breaks between postings, 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.
- 28 The example given in figure 3 shows an excerpt from a chat logfile. In this case, the postings are linearly placed one after another in the order of their arrival on the chat server. In the user chat interface, each individual posting is rendered as a block, 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
		<i>rosi's english is not the best</i>
106	Rosenstaub1979	Nö
		<i>Nope</i>
107	Rosenstaub1979	is schon zuuulang her
		<i>it's been toooooo long</i>
108	Dill	aber rosi ist prächtig
		<i>but rosi is magnificent</i>
109	Dill	prachtvoll
		<i>grand</i>
110	Rosenstaub1979	Ich glaube, so 9 Jahre
		<i>I think, about 9 years</i>
111	Rosenstaub1979	*lol* @Dill
		<i>*lol* @Dill</i>
112	Dill	9 jahre?
		<i>9 years?</i>
113	Rosenstaub1979	Ja, kommt fast hin
		<i>Yes, that's about right</i>

Figure 3: Sequence of postings in a chat room

- 29 A posting represents a category in its own right which is different from text or speech. Below, we examine the TEI elements for *divisions* and *paragraphs* (components of texts) and for *utterances* (components of spoken discourse) to check whether they would suffice to encode postings.
- 30 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, such as

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 or publisher) who makes certain structuring decisions while composing his text and, thus, divides it into a series of units (for example, according to subtopics and information units). In CMC, on the other hand, 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 in chat logfiles) or a joint achievement of the group of users (as in Wikipedia talk pages and in certain forums). Thus, the resulting structure is not based on any sort of authorial structuring 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, whereas a posting in CMC is an atomic constituent of a written dialogue determined by the ongoing dialogue’s *local* coherence.

31 For example, in figure 3, the user *Rosenstaub* sends posting 106 (“Nope”) 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, for example, 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 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, user *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*) 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.

32 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 simultaneity of verbalization, perception, and mental processing as one very central characteristic of spoken utterances is not present in postings: Due to the “pre-transmission composition” protocol discussed above, the turn-taking apparatus does not function in the same way as in spoken conversation. Postings—like texts—are first produced in their entirety; the composition process can accordingly not be tracked by the other participants, its result (after having been submitted to and transmitted by the server) can only be *read* retrospectively. In spoken conversation, on the other hand, the listeners can give immediate feedback and, thus, directly react to (and affect) the ongoing verbalization; they can anticipate the completion of turn-constructional units and negotiate turns simultaneously with the linear unfolding of the current speaker’s utterance (see, for example, Sacks, Schegloff and Jefferson 1974; Schegloff 2007).

33 Therefore, in our schema, the element <posting> is the basic structural element of a CMC document. We consider it a *macrostructural* element, but it is the pivot between the higher level macrostructural components thread and logfile (see section 3.3.2) and the *microstructure* of the content which it encloses (see section 3.5). The structure of <posting> is based on that of the existing <div> element.

34 The <div> and <posting> elements have the following similarities:

- <div> and <posting> are high-level elements, belonging to the same class (model.divLike);
- <div> and <posting> contain the major divisions of text;
- <div> and <posting> have similar internal content.

35 It is important to note that <posting>, like <div>, does not belong to the class of pLike elements. One <posting> may consist of one or more paragraphs, similar to a <div>. While a division may represent, for example, a chapter of a book, <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 <div>. In the chat example given in figure 3, all postings consist of exactly one paragraph and the portion of text exhibits no special markup, but 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.pLike element.

36 The <div> and <posting> elements have the following differences:

- <div> is a self-nesting element, while <posting> is not;
- <posting>s can only appear inside of a division which encloses one complete CMC document (such as an entire forum thread, an entire blog with user comments, or a chat logfile).

37 In other words, <posting> is a child element of <div> and shares its content model except that it does not contain divisions and does not embed itself. Normally, <posting> consists of one or more paragraphs. In some cases a posting contains a head, typically with a title.

38 Attributes in the following classes can be used with the posting element: att.ascribed, att.dateable, att.global, att.typed. The most commonly used attributes for posting are @synch and @who. @synch is used to signify the time when a posting arrives at the server. Such sequential points in time are ordered on a timeline encoded separately from the postings in the same XML document (in the <front> section, as shown in the code snippet in fig. 4 and section 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 <person> element is used for this purpose.

39 In addition, we introduce new attributes in the TEI customization specifically for use with the <posting> element: @revisedWhen, @revisedBy, and @indentLevel. The first two attributes are similar to @synch and @who but differ from them in the following aspect: they mark the time when a posting was revised and the person who revised it (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 are added to the att.ascribed class; @synch and @revisedWhen are added to the att.dateable class. The values of @synch, @who, @revisedWhen, and @revisedBy are URIs which point to a profile and to a point of a timeline. The @indentLevel attribute is added to 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 value of this attribute must be an integer from 1 to ∞ depending on the level of the indentation of the posting (see the encoding example given in fig. 5).

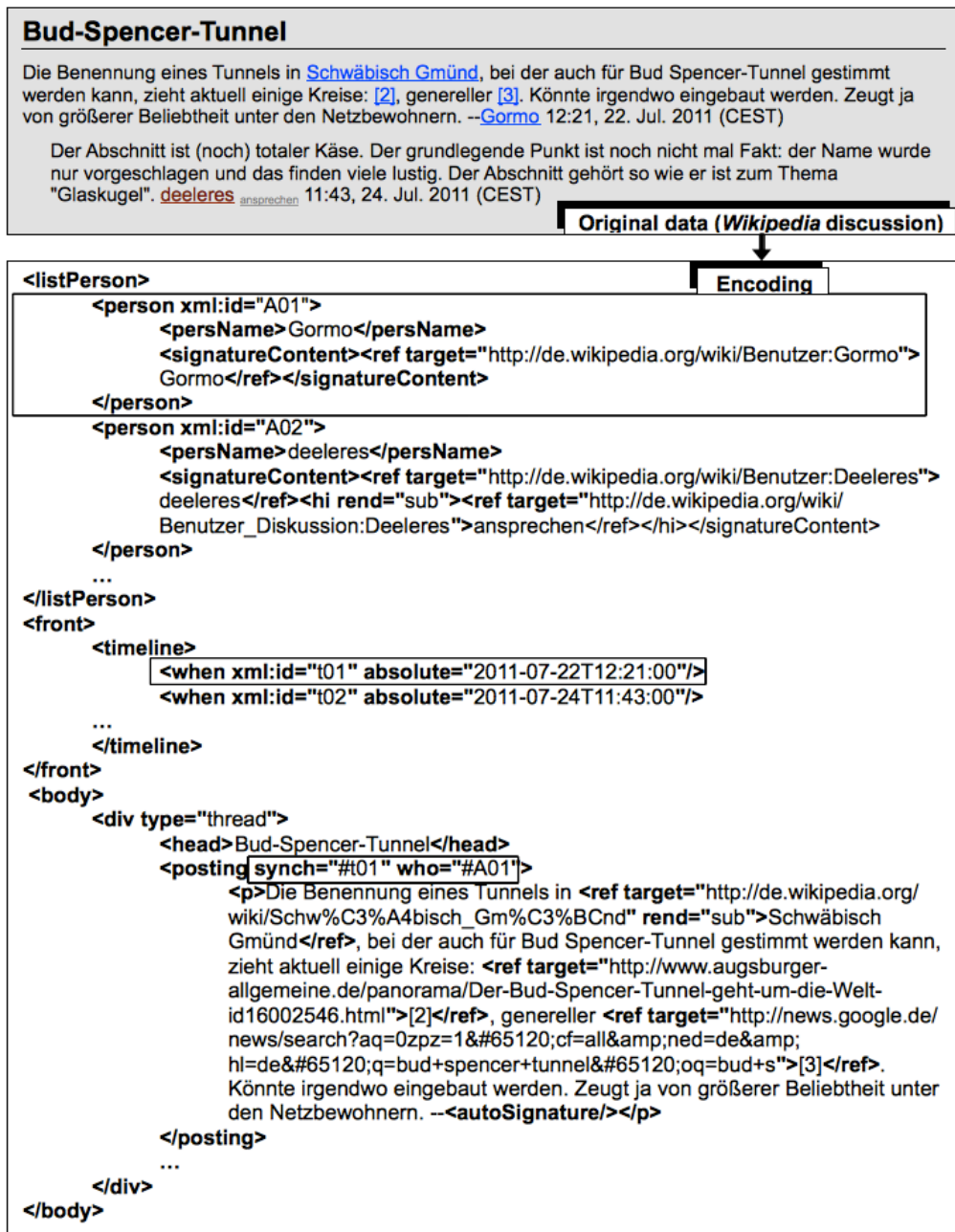


Figure 4: This example contains an encoding of a user profile, a part of the timeline, and one posting. For the complete encoding of this XML document, see <http://www.empirikom.net/bin/view/Themen/CmcTEI>.

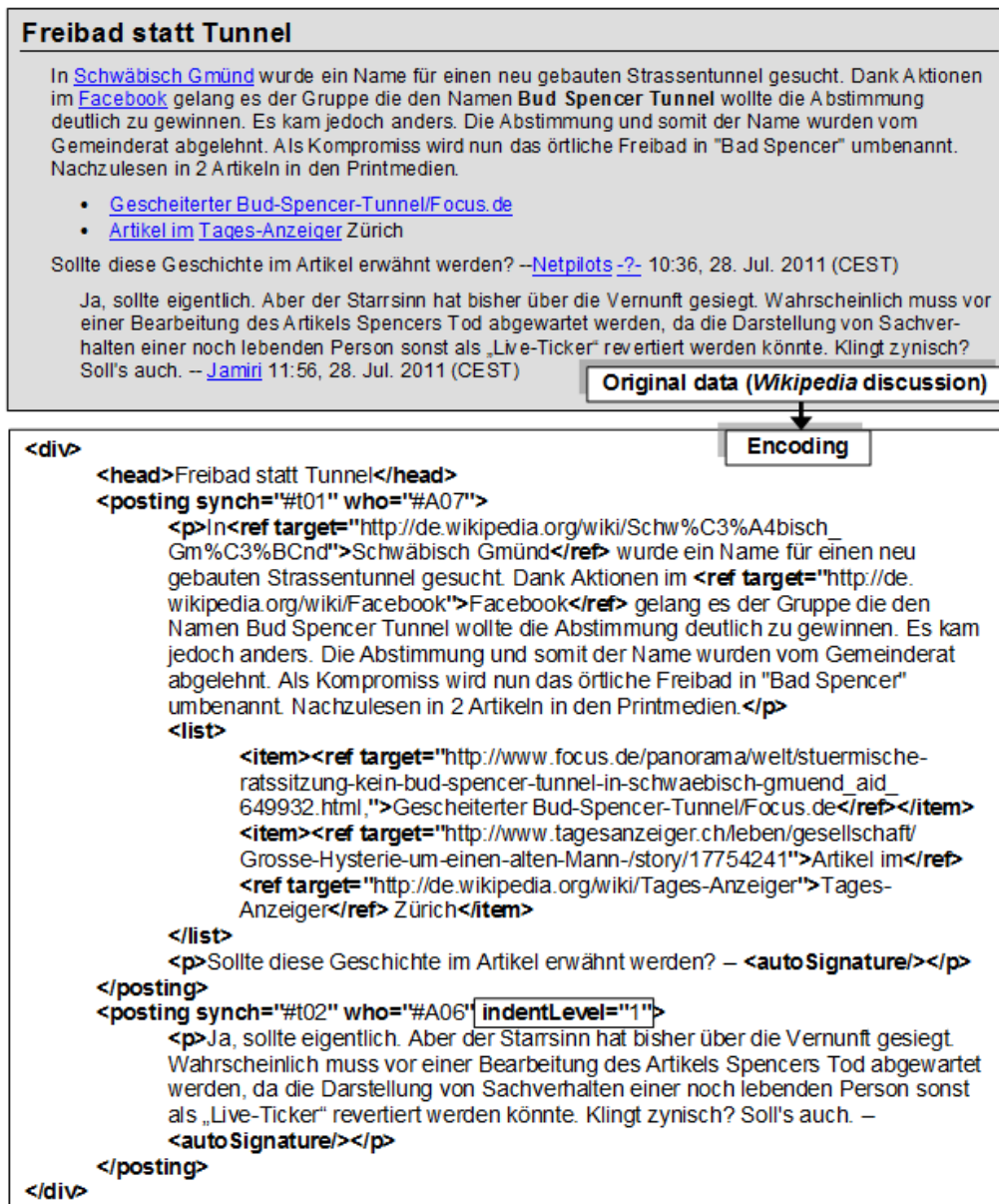


Figure 5: Encoding of postings 1 and 2 from the example given in figure 2

3.3.2. Threads and logfiles

40

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. Thus, the structuring on the macrostructure level of a CMC document has a different status from 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 `<p>` element 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 postings in chronological order based on when they reached the server (see the examples given in fig. 7)
2. *threads*, which structure the sequence of postings in two dimensions:
 - a. the above/below dimension, which usually 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 (see the example given in fig. 6).

41 To differentiate these two CMC-specific macrostructure types, we use the values *thread* and *logfile* on the @type attribute of <div>.

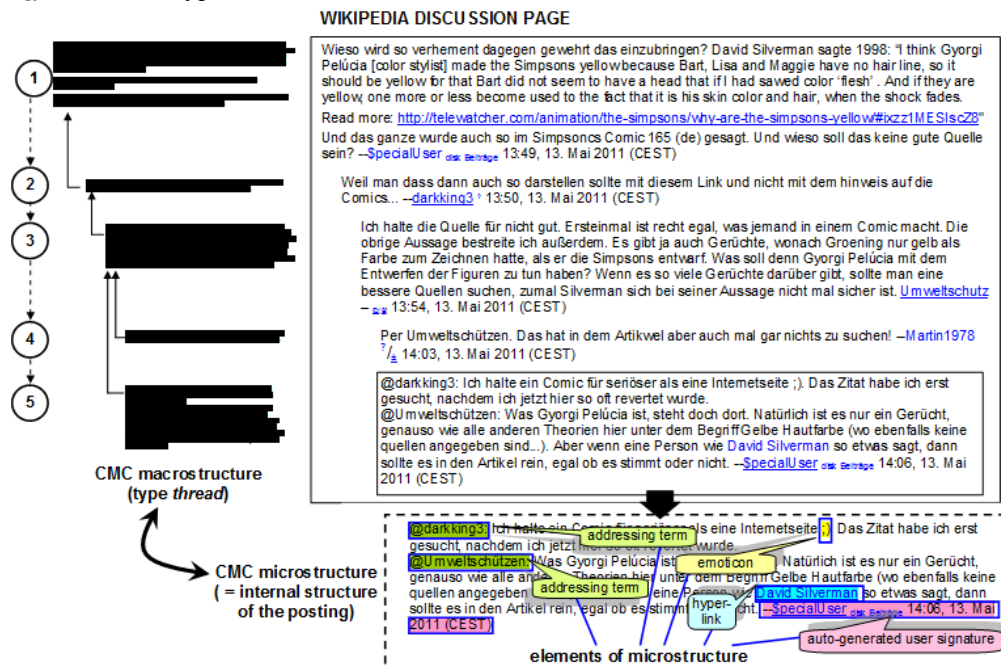


Figure 6: Differentiation between CMC macro- and microstructures in a CMC "thread" macrostructure

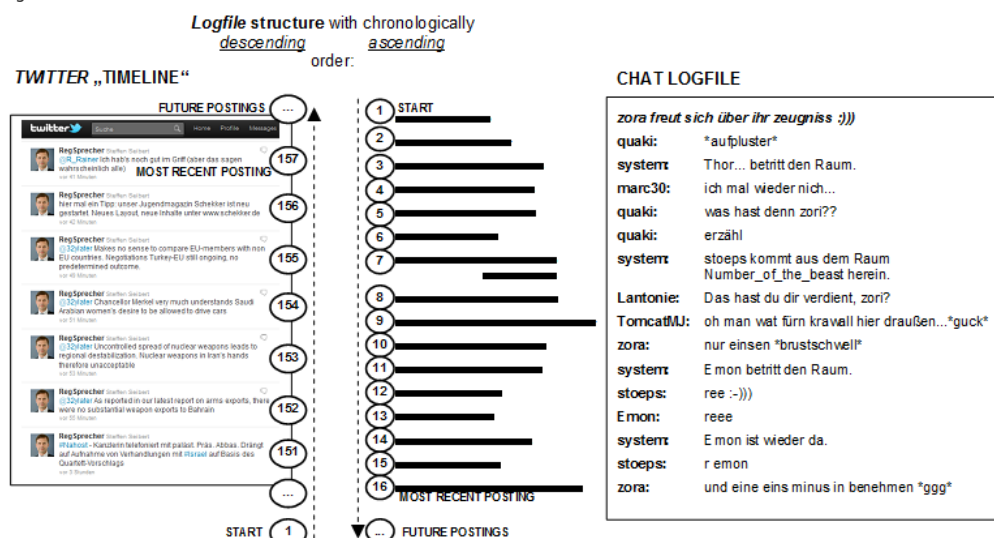


Figure 7: CMC "logfile" macrostructure

3.4. Metadata and Anonymization

3.4.1. Metadata

42 The TEI customization needs to account for metadata specific to CMC. In our context, it is convenient to add metadata to each individual document, and the TEI header is sufficient to record data relevant to the description of a CMC document. However, we want to draw the attention of the reader to the following features which are particular to the CMC document type:

1. Documents are quite difficult to identify on the Web. Mechanisms of persistent identifiers are just now gaining ground and are far from being well established. We therefore follow a double strategy: in cases where we are able to refer to a persistent identifier (as is the case with versions of Wikipedia talk pages), we include that information as a part of the source description. In cases where we cannot refer to a persistent identifier, we download the web page and store it as a digital copy and refer to it in the source description.

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 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 (postings) are situated via the @synch attribute of the element <posting> (see section 3.3.1). 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

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

44 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 in a <person> element. This profile is provided with a value of @xml:id which is unique within the particular TEI 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 linked to a person profile via the @who attribute, which points to the value of an @xml:id of a <person> element.
- Instances of user names in segments of a given posting are also linked to a <person> (see section 3.5.1.5 below).

45 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.

46 For an example of this strategy in use, see the example in figure 4 (section 3.3.1).

3.5. Elements of the Document Microstructure

3.5.1. CMC-specific Types of Interaction Signs

47 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 two—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, nor consensus about the status of these elements in a natural language grammar framework. To fill this gap, we have developed an annotation schema for these phenomena on the microstructure level of CMC documents. The basic linguistic description category of our approach is termed an *interaction sign*; in the schema, instances of interaction signs such as emoticons, acronyms, etc. are represented using the element <interactionTerm>. Below we briefly introduce the category of an interaction sign and embed it into a broader grammatical framework. By means of examples, we describe how the category and its subcategories are used for the annotation of our German reference corpus.

48 First and foremost, our schema serves the annotation needs of the DeRiK project. Some of the subcategories may be specific to German CMC, so it is clear that the annotation schema suggested below has to be developed further and discussed within the CMC community. For example, 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 elements of Internet jargon in CMC corpora. The schema favors a grammatical perspective, but it is open for extensions motivated by other fields of research such as 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”, and “wow”. Grammar frameworks usually categorize them as *interjections* (see, for example, Greenbaum 1996; McArthur et al. 1998; Blake 2008) or *Interjektionen* (DUDEN 2005), *inserts* (Biber et al. 1999; Biber et al. 2002), *discourse markers* (Schiffrin 1986), *discourse particles*, or *Gesprächspartikeln* (DUDEN 1995). These interjections are different from responsives like “yes” and “no”, which can occur in both 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, and Strecker 1997, henceforth *GDS*),¹⁰ both interjections and responsives are categorized as *Interaktive Einheiten* (henceforth *IE*). 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.¹¹ One important syntactic feature of IE is that they are not integrated in the sentence’s syntactic structure (Ehlich 1986; Trabant 1998). Instead, they are often either used as sentence-equivalent utterances (like “nö” in posting 106 of the example given in fig. 3 above) or used in front of or after the sentence boundaries (like “ja, sollte eigentlich” in posting 2 of the example given in fig. 2).

Many CMC-specific elements like emoticons and acronyms occur in the same positions and have similar functions as IEs in spoken discourse. It is, thus, not surprising that grammars—if they describe them at all—classify these elements as interjections.¹² In the STTS tagset, a standard for German part-of-speech classification,¹³ most IEs would best be annotated using the POS-Tag ITJs (*Interjektio*) or PTKANT (*Antwortpartikel*); in the CLAWS2 tagset for English,¹⁴ they would fit into the category UH (*interjection*).

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 language-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*” (derived from the English “smile”) are often used as synonyms. All these elements share topological and functional features with natural language interjections in spoken discourse. By subsuming all of these elements of Internet jargon under one category, “interaction sign”, we want to account for their functional and semantic similarities (see fig. 8).

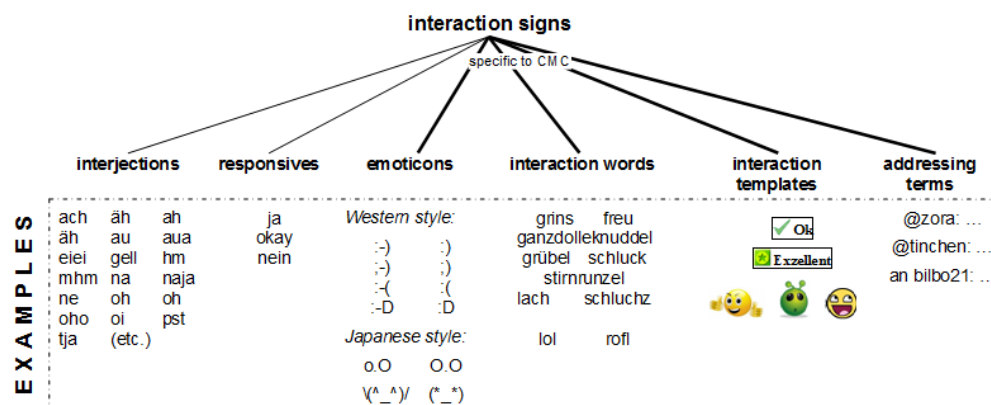


Figure 8: Typology of interaction signs (with examples)

In our schema, we introduce an element <interactionTerm> as a phrase-level element (in the model.phrase class) which encloses one or more instances of subclasses of interaction signs. The <interactionTerm> element can have members of att.global as attributes. In addition, we introduce elements for the following subclasses of interaction signs: the two subclasses

of “Interaktive Einheiten” as described by the GDS (*interjection* and *responsive*) and 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 sub-classified 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.

3.5.1.2. Emoticons

Emoticons are iconic units created using the keyboard. They are often used to portray facial expressions, and 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, for example, Japanese and Korean style variants. Postings 3 and 5 in the example given in figure 2 include Japanese-style emoticons (“Kawaiiicons”); Western-style emoticons can be found in the example given in figure 9.

- 1) Was noch fehlt ist die heutige Nutzung der Kirche. Durchlesen werde ich es mir noch, dauert nur noch ein wenig (ich denke aber, daß ich es heute noch schaffen werde!) –Grüße aus Memmingen 13:30, 25. Feb. 2009 (CET)
What is still missing is today's use of the church. I will still read all the way through it, but it will take a while longer (but I think that I will get to it today!) –Grüße aus Memmingen 13:30, 25 Feb. 2009 (CET)
- 2) Die Nutzung ist gleich im zweiten Satz erwähnt. Die Pfarrstelle ist zur Zeit unbesetzt, aber dies ist wohl kaum relevant. –Alma 13:48, 25. Feb. 2009 (CET)
The use is mentioned right off in the second sentence. The rectorate is not filled at the moment, but this is hardly relevant. –Alma 13:48, 25 Feb. 2009 (CET)
- 3) Leider nicht wirklich :o). Mach doch am besten nen Extra Absatz ganz am Schluß des Artikels == Nutzung ==. Da kommt dann rein, ob Gottesdienste stattfinden (und wann i. d. R., also z. B. Sonntags), ob Orgel/Kirchenkonzerte in dem Kirchenraum stattfinden, etc.. –Grüße aus Memmingen 15:04, 25. Feb. 2009 (CET)
Unfortunately not really ;o). The best way would be to add an extra paragraph at the end of the article ==Use==. One would write there whether mass takes place (and when normally, i.e. Sundays), whether organ/church concerts take place in the church space, etc.. –Grüße aus Memmingen 15:04, 25 Feb. 2009 (CET)
- 4) Na das ist kein Thema mache ich. –Alma 15:05, 25. Feb. 2009 (CET)
That's no problem, I'll do it. –Alma 15:05, 25 Feb. 2009 (CET)
- 5) Supil :o) *freu* etc. *g* –Grüße aus Memmingen 15:06, 25. Feb. 2009 (CET)
*Great! :o) *happy*, etc. *g* –Grüße aus Memmingen 15:06, 25 Feb. 2009 (CET)*
- 6) Orgel: Irgendwas passt da nicht in meinen Kontext: Sie wurde auf der 1899 errichten Empore aufgebaut. Sie war 1895 eine der ersten drei Hochdruckstimmenorgeln Wie soll das gehen? 1895 war vor dem Bau der Orgel...auch wäre hier die Disposition noch recht nett :o) –Grüße aus Memmingen 15:09, 25. Feb. 2009 (CET)
Organ: Something here does not fit in the context: "It was built in the gallery which was constructed in 1899. In 1895 it was one of the first three organs with high-pressure tones How can that be? 1895 was before the construction of the organ...here the arrangement would also be nice;o) –Grüße aus Memmingen 15:09, 25 Feb. 2009 (CET)

Figure 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 types of interaction signs.

In our schema, instances of emoticons are represented using 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 characters (dot, comma, colon, and the like), the resulting figure is comparable in its semiotic status to graphic characters. While some smiley faces have been included in Unicode, the variety of emoticons is still larger than can be captured by Unicode characters alone. That is why we place the <emoticon> element in the class of gLike elements.

The <emoticon> element includes attributes from the att.global class and a number of new attributes from other classes, such as @style, @systemicFunction, @contextFunction, and @topology, the first three of which are members of the att.typed class. The @style attribute describes the native region of an emoticon. The value list of

@style is currently set to *Western*, *Japanese*, *Korean*, and *Other*. The attributes @systemicFunction and @contextFunction (explained below) share the following list of values: *emotionMarker:positive*, *emotionMarker:negative*, *emotionMarker:neutral*, *emotionMarker:unspec*, *responsive*, *ironyMarker*, *illocutionMarker*, *virtualEvent*.

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. Löbner 2002). The idea is that, comparable to other lexemes, these types of emoticons (and other interaction words; see section 3.5.2.2) commonly used in CMC can be assigned a general, context-independent meaning. On the Web, there are many lists displaying the “most common emoticons” with descriptions of their 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 fig. 10), are used for marking irony. The context function of these elements in (b), thus, differs from their systemic function. On the other hand, in (a) in figure 11, the context function of “:)” is identical with the systemic function; here, the emoticon is used for displaying a positive emotion of happiness.

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*.

Icon	Meaning
>:] :-) :) :o) :] :3 :c) :> =] 8) =) :} :^)	Smiley or happy face [...]
>:D :-D :D 8-D 8D x-D xD X-D XD =-D =D =-3 =3 8-)	Laughing, big grin, laugh with spectacles
:-))	Very happy
>:[:-(:(- :c :c :-< :< :-[:[:{ >.> <.< >.<	Frown, sad
:-ll	Angry
>:] ;-) ;) *-) *) ;-) ;] ;D ;^)	Wink, smirk
>:P :-P :P X-P x-p xp XP :-p :p =p :-P :P :-b :b	Tongue sticking out, cheeky/playful [...]

Figure 10: Excerpt from the list of Western emoticons as given in the English Wikipedia, page “List of emoticons” (as of 2012-02-01)

11a:	178	system	Shadok kommt aus dem Raum Alshain herein.
			<i>Shadok comes in from the room Alshain.</i>
	185	marc30	Holla Shaddy :)
			<i>Hey Shaddy :)</i>
	189	Shadok	heya marc30 ;o)
			<i>hey marc30 ;o)</i>
11b:	536	Thor	Thor... ärgert sich immer noch, daß die franzosen den pott nicht behalten haben *gg*
			<i>Thor... is still upset that the french didn't hold on to the pott</i> *gg*



	544	Erdbeere\$	Erdbeere\$ ärgert sich mit der pott geht an frankreich und wir bekommen die küste
			<i>Erdbeere\$ feels your pain the pott goes to france and we get the coast</i>
	554	Bochum	Bochum tritt erdbeere in den arsch :-))
			<i>Bochum kicks erdbeere in the butt :-))</i>
	564	Erdbeere\$	ohh wie nett :)
			<i>ohh how nice :)</i>

Figure 11: Convergence (11a) and divergence (11b) of systemic function and context function (excerpt from document no. 2221006 in the Dortmund Chat Corpus).

3.5.1.3. Interaction Words

60

Interaction words are *symbolic* linguistic units. Their morphologic construction is based on a word or a phrase of a given language which describes expressions, gestures, bodily actions, or virtual events—for example, the units *sing*, *g* (< *grins*, “grin”), *fg* (< *fat grin*), *s* (< *smile*), *wildsei* (“being wild”) in figure 12 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):

858	Turnschuh	OHNE DEUTSCHLAND FAHRN WIR ZUR EM!
		<i>WE ARE GOING TO THE EUROPEAN CUP WITHOUT GERMANY</i>
859	system	Ryo hat die Farbe gewechselt
		<i>Ryo changed colors</i>
860	Gangrulez	jo schade
		<i>yep too bad</i>
861	system	Windy123 geht in einen anderen Raum: Forum
		<i>Windy123 is going to another room: Forum</i>
862	juliana	alle leute müssen ihre fernseher bei media markt bezahlen
		<i>all the people have to pay for their TV at media markt</i>
863	juliana	haha
		<i>haha</i>
864	Turnschuh	Es gab mal ein Rudi Völler.....es gab mal ein Rudi Völler..... 
		<i>There once was a Rudi Völler.....there once was a Rudi Völler.....</i>
865	Ryo	
		<i>*g*</i>

866	Gangrulez	hehe..das wurd eh gerichtlich gestoppt juliana
		<i>hehe..that was stopped by the courts anyway juliana</i>
867	juliana	echt?
		<i>really?</i>
868	oz	[gang:] echt ??
		<i>gang: really ??</i>
869	Gangrulez	ja
		<i>yeah</i>
870	juliana	wieso?
		<i>why?</i>
871	Gangrulez	wettbewerbsverzerrung
		<i>distortion of competition</i>
872	Naturkonstantler	Fussball ist sooo unendlich unwichtig...
		<i>Soccer is sooo incredibly unimportant...</i>
873	juliana	versteh ich nicht. ich fand es war ein cooler trick
		<i>I don't understand. I thought it was a cool trick</i>
874	Gangrulez	aber es war eine Art Glücksspiel
		<i>but it was a kind of gamble</i>
875	Turnschuh	mag auch keinen Fussball.....nur wollte ich das letzte Deutschlandspiel sehen [*fg*]
		<i>Turnschuh also doesn't like soccer.....but I would have liked to have seen the last Germany game *fg*</i>
876	Chris-Redfield	[*s*] aber net erlaubt [@ juli]
		<i>*s* but not allowed @ juli</i>
877	juliana	fußball ist nen dreck wichtig. es ist ein spiel. hauptsache, die jungen männer haben sich fitgehalten und ihrer gesundheit was getan :)
		<i>soccer isn't worth it. it's a game. Main thing, the young men have kept fit and done something for their health :)</i>
878	Gangrulez	und das entsprich nicht dem Handel [*g]
		<i>and that wasn't the deal *g</i>
879	juliana	chris, du weißt doch, daß ich ein gesetzesbrecher bin [*g*]
		<i>chris, you do know that i am a law breaker *g*</i>
880	Chris-Redfield	ja ich weiß [*s*]

		<i>yes i know *s*</i>
881	juliana	*wildsei*
		<i>*being wild*</i>
882	juliana	naja... äh.
		<i>oh well... um.</i>
883	Gangrulez	ach ich muss ja noch ne mail schreiben..
		<i>oh i have to write an e-mail..</i>
884	juliana	ich geh zu meinem buch und...
		<i>I'm going to go to my book and...</i>
885	system	Gangrulez geht in einen anderen Raum: sphere
		<i>Gangrulez goes to another room: sphere</i>
886	Naturkonstantler	vielleicht können wir ja mal eine Greencard für potentielle Fussballspieler einführen... ich werde eine Petition beim B-tag einreichen... Ja, so bin ich, ich Sorge mich um das Wohl der Allgemeinheit! *g*
		<i>maybe 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! *g*</i>
887	juliana	mal schaun
		<i>we'll see</i>
888	system	juliana verlässt den Raum
		<i>juliana leaves the room</i>

Figure 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)

61

The element `<interactionWord>` in our schema is a member of `model.global.spoken`. It shares properties of the `<kinesic>`, `<incident>`, and `<vocal>` elements in TEI. The element `<interactionWord>` is provided with attributes from the class `att.global` and several new attributes: `@formType`, `@systemicFunction`, `@contextFunction`, `@topology`, and `@semioticSource`. The attributes `@systemicFunction`, `@contextFunction`, and `@topology` are used for the `<emoticon>` element. `@formType` is in the `att.typed` class of attributes and is used to describe morphological properties of the `<interactionWord>`. The list of values is currently set to *simple*, *complex*, and *abbreviated*. The attribute `@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* (such as for *grins* “grin” and *stirnrunzel* “frown”), *gesture* (such as for *kopfschüttel* “shake head” and *wink* “wave”), *bodilyReaction* (such as for *schluck* “gulp”, *seufz* “sigh”, and *hüstel* “little cough”), *sound* (such as for *plätscher* “splash” and *blubb* “plop”), *action* (such as for *tanz* “dancing”, *knuddel* “cuddling”, *erklär* “explaining”, and *mampf* “munching”), *sentiment* (such as for *freu* “happy”), *process* (such as for *träum* “dreaming”), and *emotion* (such as for *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 :)

Original data (chat logfile)

↓
Encoding

```

<posting synch="#t536" who="#A01" >
  <p>Thor... ärgert sich immer noch, daß die franzosen den pott nicht behalten haben
    <interactionTerm>
      <interactionWord formType="abbreviated" systemicFunction="ironyMarker"
        contextFunction="ironyMarker" semioticSource="mimic" topology="back_position">
        *gg*</interactionWord>
      </interactionTerm>
    </p>
</posting>
<posting synch="#t544" who="#A02">
  <p>Erdbeere$ ärgert sich mit .... der pott geht an frankreich und wir bekommen die küste</p>
</posting>
<posting synch="#t554" who="#A03">
  <p>Bochum tritt erdbeere in den arsch
    <interactionTerm>
      <emoticon style="Western" systemicFunction="emotionMarker:positive"
        contextFunction="ironyMarker" topology="back_position">:-))</emoticon>
    </interactionTerm>
  </p>
</posting>
<posting synch="#t564" who="#A02">
  <p>
    <interactionTerm>
      <interjection>ohh</interjection>
    </interactionTerm>
    wie nett
    <interactionTerm>
      <emoticon style="Western" systemicFunction="emotionMarker:positive"
        contextFunction="ironyMarker" topology="back_position">:)</emoticon>
    </interactionTerm>
  </p>
</posting>

```

Figure 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 automatically inserts a previously prepared text or graphical element into a space of the user's choice.

The category of *interaction templates* includes *graphic smileys*, chosen by the user of a CMC environment from a finite list of elements. These often portray facial expressions but can depict almost anything; in the case of animated GIFs, they can even portray entire scenes as moving pictures. This clearly goes beyond what can be expressed using only keyboard-generated emoticons. On the other hand, users can invent new emoticons by combining keyboard characters, 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 has two possible values: *static* and *animated*.

The attributes `@systemicFunction` and `@contextFunction` have already been introduced in section 3.5.1.2, but one additional value of attribute `@systemicFunction` should be mentioned: "evaluation" is used to express whether the enclosed graphic element expresses appreciation or disapproval.

3.5.1.5. Addressing Terms

Addressing terms address an utterance to a particular interlocutor (see the examples in the postings 868 and 876 in fig. 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. While this element usually uses no attributes, our customization includes the att.global attributes. The content of <addressingTerm> is restricted to two elements: <addressMarker> and <addressee>.

The <addressMarker> element belongs to the class model.labelLike (used to gloss or explain parts of a document) and is provided with the att.global class of attributes. 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 (English: “to”; German: “an” or “für”).

The element <addressee> is placed in the model.nameLike.agent class. It includes the @who, @scope, and @formType attributes, plus those from the att.global class. Names of addressees are often addressed using abbreviated or nickname forms of their usernames, so the name of the addressee given in the addressing term might not be identical with the username of the interlocutor. We would like to enable the users of our corpus to retrieve the alternative form from the data even after the corpus data have been anonymized (as explained in section 3.4). We use the @formType attribute for this purpose and assign it the following set of values: *persNameFull*, *persNameAbbreviation*, and *persNameNickname*. Thus, the attribute @formType allows us to describe cases like the ones illustrated through the examples in figure 14:

14a:		
306	Lantonie	Lantonie heiratet Thor.... <i>Lantonie is marrying Thor....</i>
308	Lantonie	:)) :))
323	zora	wos? *eifersüchtel*@ [lanto] <i>what? *jealous*@[lanto]</i>
14b:		
104	Chris-Redfield	tom ram ist doch nicht alles im leben *g* <i>tom ram is not all there is in life *g*</i>
108	TomcatMJ	nö, aber hilft dem server weiter@ [c-r] :-) <i>no, but helps the server@[c-r] :-)</i>
14c:		
117	Raebchen	Raebchen rät allen Pärchen, nicht auf Deck zu knutschen (sowas hat die Titanic sinken lassen! habe ich im Film gesehen) <i>Raebchen advises all couples not to make out on deck (that's what made the Titanis sink! i saw it in the movie)</i>
123	McMike	*lol*@ [Raebby] *lol*@ [Raebby]
14d:		

89	McMike	könntet Ihr mich bitte zum <u>Käpten</u> ernennen?
		<i>could you all please appoint me captain?</i>
94	ineli26	ineli26 ernennt McMike zum Kapitaen
		<i>Ineli26 appoints McMike captain</i>
[...]		
160	McMike	Monk, kannst Du das steuer übernehmen?
		<i>Monk, can you take over the wheel?</i>
164	Monk	klar wohin solls gehen?
		<i>of course where to?</i>
169	McMike	Monk immer dem Fön nach
		<i>Monk keep following the Foen</i>
172	ineli26	lol @ <u>kapitaen</u>
		lol @ <u>kapitaen</u>

Figure 14: Types of addressees' names in addressing terms: abbreviated form (14a and 14b) and nickname form (14c and 14d) (excerpts from documents no. 2221006, 2221007, and 2221001 in the Dortmund Chat Corpus)

The @scope attribute is added to the att.scoping class. This attribute is used to specify whether one or more persons or groups are addressed; the values of this attribute are *all*, *group*, *individual*, and *unspec*.

The @who attribute is supposed to mark the name of the addressee (the recipient of the posting). Its value points to the value of @xml:id of the <person> element for the addressee.¹⁵

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

868	oz	gang: echt ??	
876	Chris-Redfield	*s* aber net erlaubt @ juli	Original data (chat logfile)

↓ Encoding

```

<posting synch="#868" who="#A01">
  <p>
    <interactionTerm>
      <addressingTerm>
        <addressee formType="persNameAbbreviation" who="#A07"
          scope="individual">gang:</addressee>
      </addressingTerm>
    </interactionTerm>
    echt ??
  </p>
</posting>
<posting synch="#876" who="#A02">
  <p>
    <interactionTerm>
      <interactionWord formType="abbreviated"
        systemicFunction="emotionMarker:positive" contextFunction="responsive"
        semioticSource="mimic" topology="front_position">*s*</interactionWord>
    </interactionTerm>
    aber net erlaubt
    <interactionTerm>
      <addressingTerm>
        <addressMarker>@</addressMarker>
        <addressee formType="persNameAbbreviation" who="#A10"
          scope="individual">juli</addressee>
      </addressingTerm>
    </interactionTerm>
  </p>
</posting>

```


Figure 15: Encoding snippet for postings 868 and 876 from the example in figure 12

3.5.2. User Signatures

An important element of the microstructure in postings in forums, bulletin boards, and wiki discussions is the signature text predefined by a user and inserted into a posting automatically (usually at its end). It often includes the name of the user plus additional text (such as sayings, proverbs, quotes, or personal information about the user) or 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 <person> element.

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 <autoSignature> is an empty element 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 <person> element; it is placed in the model.persStateLike class and referenced by the @target attribute on <autoSignature>.

3.5.3. Postscripts, Openers, and Closers

Some elements in CMC discourse are similar to elements used in epistolary correspondence. However, their use is less restricted than with their functional equivalents in written letters.

One element of this type is the <postscript>. In CMC, a complete posting can be marked by a user as a postscript (for example by introducing it with "p.s."); in other cases, a postscript can be a part of a paragraph (see the examples given in fig. 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>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 it still go wrong. [...]</i>
(Follow-up message of <i>user1</i> to his own prior posting in a blog discussion; anonymized)
16b:
Die genannten Quellen sind für die Fragestellung in keinsten Weise reputabel, d.h. auch danach läge Theoriefindung vor. In Volkach heisst die Mainbrücke auch nur Mainbrücke, weil es für Einheimischen nur diese eine gibt. Aber der Eigentümer, das Land Bayern, hat natürlich mehrere Mainbrücken, daher ist es nun einmal die <u>Mainbrücke Volkach</u> . Also Fahrradbrücke wird das Bauwerk sicher nicht heissen, man müsste halt mal bei der Bauverwaltung der Stadt Konstanz nachfragen. Anderenfalls dann doch gemäß reputabler Literatur auf <u>Geh- und Radwegbrücke über den Seerhein bei Konstanz</u> verschieben. -- Störfix 21:55, 13. Jul. 2011 (CEST) <u>[P.S. oder die Brücke endlich z.B. nach einem verdienten OB benennen ;-)]</u>
<i>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 <u>Main Bridge Volkach</u>. 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, <u>Geh- und Radwegbrücke über den Seerhein bei Konstanz</u>. --Störfix 21:55, 13. Jul. 2011 (CEST) <u>[P.S. or finally name the bridge after a deserving mayor ;-)]</u></i>
(Wikipedia talk page for the article "Geh- und Radwegbrücke über den Seerhein bei Konstanz")

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

CMC communication is characterized by a less conventional style of writing than in epistolary correspondence, which affects 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 Wikipedia talk pages) 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 necessarily the first or last part of the message (see example below). Additionally, an opener or closer element can appear more than once in a posting.

80 Unfortunately, the elements of the current TEI P5 framework which come closest to these structures (the <opener> and <closer> elements) are too restricted in their 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. For us to use these elements, the content model for <div>s would have to be loosened to allow these elements to appear in other places. Specifically, it would be useful if the <opener> and <closer> elements could join the *inter-level elements* so that they would be able to appear within as well as in between chunks of text. In the current version of our schema, we use <seg> elements for the annotation of openers and closers in CMC postings and use a @type attribute with a value of “opener” or “closer” (see the example given in fig. 17).

```
<posting who="#A02" synch="#02" indentLevel="1">
  <p><seg type="opener">Servus <persName ref="#A03"/></seg>! Kennst du die
  Bearbeitung in der neuen <ref target="http://www.efloras.org/florataxon.aspx?
  flora_id=2&#65120;taxon_id=119600">Flora of China</ref>? Zwei der drei aus
  China angegebenen Arten sind zumindest nicht allgemein akzeptiert. Grundsätzlich
  muss man aber auch bei allen Arten, die aus der alten Sowjetunion beschrieben
  wurden, vorsichtig sein: Die hatten so eine Art Dogma, dass es keine Unterarten
  geben darf. So ist halt automatisch alles, was nach einer phänotypisch
  abgrenzbaren Sippe ausgesehen hat, gleich als Art beschrieben worden.
  <seg type="closer">Grüße</seg> --<autoSignature/></p>
</posting>
```

Figure 17: Opener and closer inside one posting, encoded using the <seg> element

4. Conclusions and Outlook

81 We have shown in this paper that the TEI Guidelines offer an appropriate way of structurally encoding documents of various CMC genres. We demonstrated 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.

82 Customization of the TEI Guidelines is one way of adapting the TEI encoding framework to new genres and document types. However, considering 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. 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 (among 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 section 3.5.1). 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 (see section 3.5.2), the position of the equivalent TEI elements in the structure of the postings is less restricted than in epistolary correspondence. In cases like these, a modification of existing TEI elements (the elements <postscript>, <opener>, and <closer>) would ideally account for both CMC’s orientation toward traditional text types and text elements as well as CMC’s free and creative use and modification.

83 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 might 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 constant 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 media of CMC which will emerge in the future. We are confident that the TEI Guidelines will provide an appropriate framework for this. We hope that 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 not using the Latin alphabet).

For DeRiK in particular, we are facing the following challenges in the near future:

- *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 10 million tokens per year. We have to clear the rights of many of the text sources unless they have not already been cleared by the providers, as is the case with Wikipedia talk pages, for example. We hope that we can acquire substantial portions of data from projects focused on empirical research in the field of CMC (including 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, and more users and researchers could find and use this data easily.
- *Analyzing CMC texts linguistically:* Software for 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 and sentence boundary detection to part-of-speech tagging and syntactic parsing. We hope to have shown with the examples in this paper 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 well-formulated—texts and dialogues (such as on Wikipedia talk pages or scientific blogs) to very informal and highly speech-like contributions in some chat sessions. Tools for the linguistic analysis of CMC should be able to cover the whole range.
- *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 (such as 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 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 transform CMC data from various sources to the TEI target format; ideally this will be a framework which can be parameterized for each individual source. In addition, scripts will be needed to 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. Additional facilities will be provided by the DWDS framework (see section 2.2).

Bibliography

References

- Beißwenger, Michael. 2002. "Getippte 'Gespräche' und ihre trägermediale Bedingtheit: Zum Einfluß technischer und prozeduraler Faktoren auf die kommunikative Grundhaltung beim Chatten." In *Moderne Oralität*, edited by Ingo W. Schröder and Stéphane Voell, 265–299. Marburg: Reihe Curupira.
- . 2003. "Sprachhandlungskoordination im Chat." *Zeitschrift für germanistische Linguistik* 31 (2): 198–231.
- . 2007. *Sprachhandlungskoordination in der Chat-Kommunikation*. Linguistik, Impulse, & Tendenzen 26. Berlin: de Gruyter.
- . 2010. "Chattern unter die Finger geschaut: Formulieren und Revidieren bei der schriftlichen Verbalisierung in synchroner internetbasierter Kommunikation." In *Nähe und Distanz*, edited by Vilmos Ägel and Mathilde Hennig, 247–294. *Linguistik, Impulse, & Tendenzen* 35. Berlin: de Gruyter.
- Beißwenger, Michael and Angelika Storrer. 2011. "Digitale Sprachressourcen in Lehramtsstudiengängen: Kompetenzen – Erfahrungen – Desiderate." In *Language Resources and Technologies in E-Learning and Teaching*, edited by Frank Binder, Henning Lobin, and Harald Lünen. *Special issue, Journal for Language Technology and Computational Linguistics* 26 (1): 119–139. http://media.dwds.de/jlcl/2011_Heft1/9.pdf.
- . 2008. "Corpora of Computer-Mediated Communication." In *Corpus Linguistics. An International Handbook. Volume 1*, edited by Anke Lüdeling and Merja Kytö, 292–208. *Handbooks of Linguistics and Communication Science* 29.1. Berlin: de Gruyter.
- Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2012. "DeRiK: A German Reference Corpus of Computer-Mediated Communication." *Digital Humanities 2012*. <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/derik-a-german-reference-corpus-of-computer-mediated-communication/>.
- Biber, Douglas et al. 1999. *Longman Grammar of Spoken and Written English*. Edinburgh: Pearson Education Limited.
- Biber, Douglas, Susan Conrad and Geoffrey Leech. 2002. *Longman Student Grammar of Spoken and Written English*. Edinburgh: Pearson Education Limited.
- Blake, Barry J. 2008. *All About Language*. New York: Oxford University Press.
- Crystal, David. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.
- Danet, Brenda, and Susan C. Herring, eds. 2007. *The Multilingual Internet. Language, Culture, and Communication Online*. New York: Oxford University Press.
- December, John. 1996. "Units of Analysis for Internet Communication," *Journal of Computer-Mediated Communication* 1 (4). Accessed February 03, 2012, <http://jcmc.indiana.edu/vol1/issue4/december.html>.
- DUDEN. 1995. *Die Grammatik*. 5th ed. Mannheim: Bibliographisches Institut.
- DUDEN. 2005. *Die Grammatik*. 7th ed. Mannheim: Bibliographisches Institut.
- Ehlich, Konrad. 1986. *Interjektionen*. Tübingen: Niemeyer.
- Ferrara, Kathleen, Hans Brunner, and Greg Whitemore. 1991. "Interactive written discourse as an emergent register." *Written Communication* 8 (1): 8–34.
- Garcia, Angela Cora, and Jennifer Baker Jacobs. 1998. "The Interactional Organization of Computer Mediated Communication in the College Classroom." *Qualitative Sociology* 21 (3): 299–317.
- . 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.
- Geyken, Alexander. 2007. "The DWDS corpus: A reference corpus for the German language of the 20th century". In *Collocations and Idioms*, edited by Christiane Fellbaum, 23–40. London: Continuum Press.
- Greenbaum, Sidney. 1996. *The Oxford English Grammar*. New York: Oxford University Press.
- Herring, Susan C. 1996. "Introduction." In *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, edited by Susan C. Herring, 1–10. *Pragmatics & Beyond* n.s. 39. Amsterdam: John Benjamins.
- . 1999. "Interactional Coherence in CMC." *Journal of Computer-Mediated Communication* 4 (4). <http://jcmc.indiana.edu/vol4/issue4/herring.html>.
- Herring, Susan C., ed. 1996. *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives. Pragmatics & Beyond* n.s. 39. Amsterdam: John Benjamins.

Herring, Susan, ed. 2010/2011. *Computer-Mediated Conversation. Special issue, Language@Internet* 7/8. <http://www.languageatinternet.org/>.

Hoffmann, Ludger. 2004. "Chat und Thema." In *Internetbasierte Kommunikation*, edited by Michael Beißwenger, Ludger Hoffmann, and Angelika Storrer, 103–122. *Osnabrücker Beiträge zur Sprachtheorie* 50.

Klappenbach, Ruth, and Wolfgang Steinitz, eds. 1962–1977. *Wörterbuch der deutschen Gegenwartssprache*. 6 vols. Berlin: Akademie-Verlag.

Löbner, Sebastian. 2002. *Understanding Semantics*. London: Edward Arnold Publishers.

McArthur, Tom, ed. 1998. *Concise Oxford Companion to the English Language*. Oxford: Oxford University Press.

Ogura, Kanayo, and Kazushi Nishimoto. 2004. "Is a Face-to-Face Conversation Model Applicable to Chat Conversations?" Paper presented at the Eighth Pacific Rim International Conference on Artificial Intelligence, 2004. <http://ultimavi.arc.net.my/banana/Workshop/PRICAI2004/Final/ogura.pdf>.

Reynaert, Martin, Nelleke Oostdijk, Orphée De Clercq, Henk van den Heuvel, and Franciska de Jong. 2010. "Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus," *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*: 2693–2698. Accessed February 03, 2012 http://eprints.eemcs.utwente.nl/18001/01/LREC2010_549_Paper_SoNaR.pdf

Runkehl, Jens, Peter Schlobinski, und Torsten Siever. 1998. *Sprache und Kommunikation im Internet: Überblick und Analysen*. Opladen: Westdeutscher Verlag.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language* 50 (4): 696–735.

Schegloff, Emanuel A. 2007. *Sequence Organization in Interaction*. Vol. 1 of *A Primer in Conversation Analysis*. Cambridge: Cambridge University Press.

Schiffrin, Deborah. 1986. *Discourse markers*. Vol. 5 of *Studies in Interactional Sociolinguistics*. Cambridge: Cambridge University Press.

Schönfeldt, Juliane, and Andrea Golato. 2003. "Repair in Chats: A Conversation Analytic Approach." *Research on Language and Social Interaction* 36 (3): 241–284.

Storrer, Angelika. 2001. "Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation." In *Sprache im Alltag: Beiträge zu neuen Perspektiven in der Linguistik; Herbert Ernst Wiegand zum 65. Geburtstag gewidmet*, edited by Andrea Lehr, Matthias Kammerer, Klaus-Peter Konerding, Angelika Storrer, Caja Thimm, and Werner Wolski, 439–465. Berlin: de Gruyter.

———. 2009. "Rhetorisch-stilistische Eigenschaften der Sprache des Internets." In *Rhetorik und Stilistik – Rhetorics and Stylistics: Ein internationales Handbuch historischer und systematischer Forschung*, edited by Ulla Fix, Andreas Gardt, and Joachim Knappe, 2211–2226. Berlin: de Gruyter.

TEI Consortium. 2012. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. <http://www.tei-c.org/Guidelines/P5/>.

Trabant, Jürgen. 1998. *Artikulationen: Historische Anthropologie der Sprache*. Frankfurt: Suhrkamp.

Werry, Christopher C. 1996. "Linguistic and interactional features of Internet Relay Chat." In *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, edited by Susan C. Herring, 47–63. *Pragmatics & Beyond* n.s. 39. Amsterdam: John Benjamins.

Zifonun, Gisela, Ludger Hoffmann, and Bruno Strecker. 1997. *Grammatik der deutschen Sprache*. 3 vols. *Schriften des Instituts für deutsche Sprache* 7.1–7.3. Berlin: de Gruyter.

Zitzen, Michaela, and Dieter Stein. 2005. "Chat and conversation: a case of transmedial stability?" *Linguistics* 42 (5): 983–1021.

WWW Resources

ARD/ZDF Onlinestudie (1997–2011). <http://www.ard-zdf-onlinestudie.de/>.

Digitales Wörterbuch der deutschen Sprache (DWDS). <http://www.dwds.de/>.

Dortmunder Chat-Korpus. <http://www.chatkorpus.tu-dortmund.de/>.

Grammis 2.0: das grammatische Informationssystem des Instituts für deutsche Sprache (IDS). <http://hypermedia.ids-mannheim.de/>.

"Online documentation of the DeRiK TEI schema for the representation of computer-mediated communication." <http://www.empirikom.net/bin/view/Themen/CmcTEI>.

“Projekt: Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK).” <http://www.empirikom.net/bin/view/Themen/DeRiK>.

Scientific network (DFG). “Empirische Erforschung internetbasierter Kommunikation“ (“Empirical Research on Internet-based Communication”). <http://www.empirikom.net>.

“STTS Tag Table.” Institute for Natural Language Processing. <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>.

Text Encoding Initiative (TEI). <http://www.tei-c.org/index.xml>.

“UCREL CLAWS2 Tagset.” University Centre for Computer Corpus Research on Language. <http://ucrel.lancs.ac.uk/claws2tags.html>.

Notes

1 <http://www.ard-zdf-onlinestudie.de>

2 For a brief description of the project, see also <http://www.empirikom.net/bin/view/Themen/DeRiK>.

3 <http://www.dwds.de/>

4 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.

5 <http://www.empirikom.net/bin/view/Themen/CmcTEI>

6 <http://www.chatkorpus.tu-dortmund.de>

7 <http://www.empirikom.net/bin/view/Themen/DeRiK>

8 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.

9 Recent overviews are given in Storrer 2009 and Herring 2010/2011.

10 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.

11 See 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”.

12 See, for example, DUDEN (2005, sec. 892) and Ehlich (1986).

13 See the STTS tag table: <http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html>.

14 See the CLAWS2 tagset: <http://ucrel.lancs.ac.uk/claws2tags.html>.

15 This is part of the anonymization strategy discussed in section 3.4.

Cite this article

Electronic reference

Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer and Angelika Storrer, « A TEI Schema for the Representation of Computer-mediated Communication », *Journal of the Text Encoding Initiative* [Online], Issue 3 | November 2012, Online since 15 October 2012, connection on 05 November 2012. URL : <http://jtei.revues.org/476> ; DOI : 10.4000/jtei.476

Authors

Michael Beißwenger

Michael Beißwenger is a researcher and lecturer for German Linguistics at TU Dortmund University. He graduated from the University of Heidelberg with an M.A. in German Philology and History (2000) and finished his Ph.D. (“Dr.phil.”) in German Linguistics at TU Dortmund University with a monograph on interactional management in chats (“Sprachhandlungskoordination in der

Chat-Kommunikation,” Berlin/New York: de Gruyter, 2007). Since 2010, he is the coordinator of the scientific network “Empirical Research on Internet-based Communication” (<http://www.empirikom.net/>) funded by the German Research Foundation (DFG).

Maria Ermakova

Maria Ermakova is studying Historical Linguistics (M.A.) at Humboldt University Berlin. Since 2010 she has been working as research assistant for the Digital Dictionary of the German Language project (DWDS) at the Berlin-Brandenburg Academy of Sciences (BBAW).

Alexander Geyken

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

Lothar Lemnitzer

Lothar Lemnitzer is a lexicographer and researcher at the Berlin-Brandenburg Academy of Sciences (BBAW). He has written introductory books in German about corpus linguistics and lexicography. He graduated from the University of Heidelberg and finished his Ph.D. (“Dr. phil.”) in English Linguistics at the University of Münster. He currently uses large corpora of contemporary German as a basis for the compilation of articles for the Digital Dictionary of German language (DWDS).

Angelika Storrer

Angelika Storrer is professor for German linguistics at TU Dortmund University since 2002. Her research interests include computational lexicography, corpus-based methods in linguistics, and language on the Internet. As a member of the Berlin-Brandenburg Academy of Sciences (BBAW) she is involved in the work on the Digital Dictionary of German language (DWDS).

Copyright

TEI Consortium 2012 (Creative Commons Attribution-NoDerivs 3.0 Unported License)

Abstract

The paper presents an XML schema for the representation of genres of computer-mediated communication (CMC) that 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 those units regarded as “typical” for language use on the Internet. The schema introduces an element <posting>, which describes stretches of text that are sent to the server by a user at a certain point in time. Postings are the main constituting elements of *threads* and *logfiles*, which, in our schema, are the two main types of CMC macrostructures. For the microlevel of CMC documents (that is, 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 easy anonymization of CMC data for purposes in which the annotated data are 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.

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 that 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 future new version of the TEI Guidelines.

Index terms

Keywords : computer-mediated communication, CMC, web genres, thread, logfile, forum, chat