Proposal for a mini panel:
Towards an encoding standard for social media and CMC: Experiences from German and French corpus projects using TEI

Submission for *ird-cmc-rennes: Social Media and CMC Corpora for the eHumanities*. International Research Days, 23–24 October, 2015, Rennes (France)

**Format of this submission:** Our proposal of a mini panel includes two papers (Beißwenger et al.) and (Chanier et al.). If accepted, we would like to introduce the panel with a little introduction (10-15 minutes) to the basics of text encoding with the TEI framework and some general challenges in modeling CMC with TEI. We would then present and discuss the two papers (= 40 minutes presentation + 20 minutes discussion). All in all, the mini panel would last 75 minutes.

**Motivation for this panel**

The internet and social media have given rise to a broad range of new communicative genres such as chats, forums, text messaging (SMS, WhatsApp), interactions on wiki talk pages and in blog comments, via Twitter, on social network sites, and in multimodal 3D environments. A standard for the representation of these genres which is compliant with existing standards for text and speech corpora in the field of Digital Humanities would foster interoperability between language resources as well as the analysis and automatic exploitation of resources of that kind in several respects:

- It would allow scholars to build interoperable CMC corpora for different languages and thus enhance the empirical basis for doing *CMC research across languages and cultures.*

- It would allow scholars to build CMC resources which are interoperable with existing text and speech corpora and thus pave the way for corpus-based *research on language use across different types of corpora* (= comparative analysis of language use in CMC, in edited text and in spoken language).

- Through including models for the description of not only verbal but also of non-verbal acts, it would allow scholars to describe and analyze CMC *across different modalities.*

Since 2013, corpus projects from several European countries and for different languages have been cooperating in a special interest group (SIG) of the *Text Encoding Initiative (TEI, http://tei-c.org)* on the creation of a standard for encoding CMC and social media genres. As a de-facto standard for the representation of texts in digital form, the TEI encoding framework (TEI P5) provides the opportunity to benefit from the abovementioned potentials even though its current version does not include any models for the representation of the linguistic and structural peculiarities of social media and CMC genres. The SIG “computer-mediated communication” is designing such models and testing them in corpus projects and with data from a broad range of genres (cf. Beißwenger et al. 2012,
Chanier et al. 2014, Margaretha/Lüngen 2014). The goal of the SIG is to propose an addition to the TEI standard in the near future.

The two papers of this mini panel will outline intermediate results and modeling suggestions (CMC-specific models and schemas, annotation experiences) from the work of the SIG on the example of German and French CMC and social media corpora. A discussion of these suggestions with CMC researchers and representatives from other CMC corpus projects is highly welcome. The results of the discussion shall be taken into account in the further work of the SIG towards a common schema for social media and CMC.

Paper 1:
Schemas and experiences from modeling German CMC corpora in TEI

Authors: Adrien Barbaresi³, Michael Beißwenger¹, Eric Ehrhardt², Alexander Geyken³, Axel Herold³, Marc Kupietz⁴, Lothar Lemnitzer³, Harald Lüngen⁴, Angelika Storrer², Andreas Witt⁴

¹ TU Dortmund University (D) ² University of Mannheim (D) ³ Berlin-Brandenburg Academy of Sciences and the Humanities (D) ⁴ Institut für deutsche Sprache, Mannheim (D)

This paper reports about TEI schemas and modeling experiences from several German corpus projects:

- **ChatCorpus2CLARIN**: The goal of this curation project of the German CLARIN initiative (CLARIN-D, [http://www.clarin-d.de](http://www.clarin-d.de)) is to integrate an existing, annotated 1MWord chat corpus for German ([Dortmunder Chat-Korpus, http://www.chatkorpus.tu-dortmund.de](http://www.chatkorpus.tu-dortmund.de)) into the corpus infrastructures of the CLARIN-D centres at the Berlin-Brandenburg Academy of Sciences (BBAW) and of the Institut für Deutsche Sprache (IDS), Mannheim. Together with the integration, the corpus will be converted into a TEI-conformant representation format and enhanced with additional linguistic annotations (part of speech tagging, cf. the submission of Bartz et al.). The project is headed by Michael Beißwenger and Angelika Storrer and located at the University of Mannheim.

- **German Wikipedia and Usenet news corpora in DeReKo**: In 2014 and 2015, a Wikipedia corpus and a corpus of Usenet news messages have been compiled for integration in DeReKo, the German Reference Corpus ([Kupietz/Lüngen 2014](#)). For this purpose, features of the TEI for CMC proposal by Beißwenger et al. (2012) have been adapted and integrated into I5, the TEI customisation used for corpus representation in DeReKo ([Lüngen/Sperberg-McQueen 2012](#)). To represent the reference structure of news messages, further elements suggested by the TEI SIG “Correspondence” have been adapted. The discussion part of the German Wikipedia corpus (264 million tokens) and a German news corpus (20 million tokens with redundancies, and growing) have been automatically compiled and marked up according to the resulting extended version of I5 (see Margaretha/Lüngen 2014 and Schröck in preparation).
- **Deutsches Referenzkorpus zur internetbasierten Kommunikation (DeRiK):** The goal of the DeRiK project is to build a reference corpus of German CMC including data from the most prominent CMC genres. The corpus shall form a new component of a collection of corpora of written German provided by the Berlin-Brandenburg Academy of Sciences (BBAW) as part of the DWDS project (Geyken 2007, [http://www.dwds.de](http://www.dwds.de)). DeRiK is a cooperative project of TU Dortmund and the University of Mannheim with the DWDS project group at the BBAW (Beißwenger et al. 2013, Beißwenger/Lemnitzer 2013).

- **German WhatsApp Corpus:** Between November 2014 and January 2015, a consortium of linguists from seven German universities has been collecting a 2.2 MWord collection of German WhatsApp messages and dialogues (initiated and coordinated by Beat Siebenhaar, University of Leipzig; [http://www.whatsapp-deutschland.de/](http://www.whatsapp-deutschland.de/)). In 2015, this collection shall be enhanced with annotations in order to create a useful corpus for linguistic analyses of the language use in mobile messaging apps.

- **DWDS Blog corpus:** The corpus is based on specialized web archives collected using focused Web crawling techniques. The blogs feature German text under CC license (tested automatically and then verified manually) published using WordPress (to design precise content extraction tools). The corpus currently comprises 174,386 webpages for a total of 97 million tokens, it can be queried online on [zwei.dwds.de](http://zwei.dwds.de/). Comparative analyses of blog posts and comments are under way (Barbaresi/Würzner 2014).

The paper will outline the encoding architecture and basic features of the TEI schemas that have been designed for these projects and give an overview of experiences and practices in using these schemas for corpus annotation. From a linguistic point of view, a fundamental challenge in adopting the TEI for the representation of social media and CMC lies in the hybrid nature of the respective genres: Written CMC shares characteristics both with written text (= written language, use of text structuring and formatting) and spoken conversation (= dialogic, sequential organization) while at the same time it differs from both genres. Even though a range of models from the current TEI P5 standard can be adopted for the representation of structural features in CMC, written CMC cannot completely be represented using the TEI modules “text structure” and “transcriptions of speech” (cf. Beißwenger et al. 2012). By the help of examples, we will discuss for which structural and linguistic features of CMC we need new, CMC-specific models, and how these models could be implemented in the TEI encoding framework. In the current version of the schemas presented in the paper, these models are defined as customizations of the TEI standard (cf. [http://www.tei-c.org/Guidelines/Customization/](http://www.tei-c.org/Guidelines/Customization/)). In view of the increasing interest in building and analyzing CMC data in the humanities and computer sciences, a future version of the TEI guidelines should implement models for CMC as part of the standard.
The CoMeRe French CMC corpora and their modeling in TEI

Authors: Thierry Chanier1, Céline Poudat2, Ciara R. Wigham3

1 University Blaise Pascal (Fr) 2 University of Nice Sophia Antipolis (Fr), 3 Université Lumière Lyon 2 (ICAR)

CoMeRe (acronym which in French stands for network mediated communication) is a national project involving researchers from 8 different research units to develop a repository of CMC all modeled within the same extension of the TEI (Chanier et al. 2014). The project was carried out from 2013 to 2015 with the support of Corpus-Ecrits (http://corpusecrits.huma-num.fr/, a national research consortium on written corpora) and Ortolang (http://www.ortolang.fr, a national infrastructure for tools and corpora on French language).

Three key principles underlie CoMeRe: variety, openness and standards.

"Variety" is one of our keywords since we have assembled interactions stemming from networks such as the Internet or telecommunications (mobile phones), as well as mono- and multimodal, and synchronous and asynchronous communications. The genres covered within CoMeRe include text or oral chats, email, discussion forums, blogs, tweets, audio- graphic conferencing systems (conference systems with text, audio, and iconic signs for communication), or even collaborative working/learning environments with verbal and nonverbal communication. The corpus also offers a variety of discourse situations: public or private conversations, as well as informal, learning, and professional situations. The largest part of our corpus originates from existing corpora, which had been previously collected by the project members. Other parts, such as Wikipedia talk pages, have been extracted from the Web.

"Openness" is our second keyword. The first set of 11 corpora has been released (http://hdl.handle.net/11403/comere) as open data on Ortolang. This openness is mainly driven by the future inclusion of CoMeRe within the forthcoming French National Corpus - the latter is expected to become a reference for studies in French linguistics. On the other hand, our wish to release CoMeRe corpora as open data stems from the fact that, although studies on new CMC communication genres draw much attention, there is currently no existing dataset with significant coverage to form the basis for systematic research.

"Standards" refers to two different aspects. Firstly, corpora have been structured and referred to in a uniform way. The TEI-IS is the model developed as an extension of the TEI in order to encompass the Interaction Space (IS) of CMC multimodal discourse. "Standards" also refers to the uniform basic level of automatic annotations, related to segmentation and part of speech (POS) tagging which is underway. The IS model is an abstract concept temporally situated at the point when interactions occur within an online location. The online location is defined by the properties of the set of environments used by the participants. Participants in the same IS can interact. They interact through input devices mainly producing visual or oral signals.
The present communication will expose the IS model, and the classification of multimodal acts we developed to distinguish between verbal (as studied by corpus linguistics) and non-verbal acts related to body, groupware, and iconic systems (Ciekanski/Chanier 2008; Wigham/Chanier 2013a; Wigham/Chanier, 2013b). Examples of all these different types of multimodal interactions will be given out of samples extracted from our corpora.

References


Ciekanski, Maud; Chanier, Thierry (2008): Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. In: Recall, vol. 20 (2), Cambridge University Press. 162-182. DOI:10.1017/S0958344008000426


TEI-SIG “Correspondence”. Documentation and ODD. http://www.tei-c.org/Activities/SIG/Correspondence/
https://github.com/TEI-Correspondence-SIG/correspDesc

Wigham, Ciara R.; Chanier, Thierry (2013a): A study of verbal and non-verbal communication in Second Life – the ARCH121 experience, ReCall 25, 63-84. DOI:10.1017/S0958344012000250 and URI:edutice-00674138