Panel:

TEI across corpora, languages and genres: Towards a standard for the representation of social media and computer-mediated communication

Proposal for: **Text Encoding Initiative: connect, animate, innovate.** 2015 Annual Conference and Members' Meeting of the TEI Consortium, 28–31 Oct 2015, Lyon.

Panel organizers: Michael Beißwenger
TU Dortmund University (michael.beisswenger@tu-dortmund.de)
Thierry Chanier
Université Blaise Pascal, Clermont-Ferrand (thierry.chanier@univ-bpclermont.fr)

Introduction

The panel presents results and ongoing work from corpus projects in which TEI-P5 has been adopted for the representation and linguistic annotation of genres of social media and computer-mediated communication (CMC). It relates to the work of the TEI-SIG "computer-mediated communication" which is developing TEI models for the representation of CMC genres and testing these models for a broad range of genres (ranging from "text-only" genres such as chat and SMS to multimodal genres such as learning environments and Second Life) and in corpus building initiatives for various European languages.

The goal of the panel is to give an overview of models and practices in representing CMC in TEI on the example of German and French CMC corpora. A documentation and ODD files of the schemas developed by the group will be made available in the TEI wiki and be announced via the TEI mailing list before the conference so that everybody who is interested in participating in the discussion can examine the CMC models in advance.

The discussion in the panel shall serve as an opportunity for collecting feedback on these models and schema drafts from a broader community within the TEI who is interested in adapting TEI-P5 for the representation of new (digital) genres. This feedback will be taken into consideration when revising the models and – as a next step after the conference – preparing feature requests for adapting the TEI for CMC.[1]

[1] **Organizational remark:** It would be great if the panel and the meeting of the CMC-SIG could both be scheduled for Thursday 29/10/2015 (the panel in the morning and the SIG meeting in the afternoon) in order to be able to proceed with a more detailed discussion of selected issues raised during the panel at the SIG meeting.

Panel slot #1: *Paper:*

TEI across corpora, languages and genres: How TEI models will enhance the toolkit of CMC research in the Humanities (20 minutes)

Michael Beißwenger[1], Thierry Chanier[2]

[1] TU Dortmund University (D) [2] Université Blaise Pascal, Clermont-Ferrand (F)

The internet and social media have given rise to a broad range of new communicative genres which are subsumed under the term *computer-mediated communication* (*cmc*) – genres such as chats, forums, text messaging (SMS, WhatsApp), interaction on wiki talk pages and in blog comments, via Twitter, on social network sites, and in multimodal 3D environments. A TEI standard for the representation of those genres and their structural and linguistic peculiarities is a desideratum both in the fields of digital humanities and computer sciences. Such a standard would foster interoperability between language resources as well as the analysis and automatic exploitation of resources of that kind in several respect:

- It would allow scholars for building interoperable CMC corpora for different languages and thus enhance the empirical basis for doing *CMC research across languages and cultures*.
- It would allow scholars for bulding CMC resources which are interoperable with text and speech corpora that are already represented in TEI and thus pave the way for corpus-based *research on language use across different types of corpora* (= comparative analysis of the 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 analyse CMC *accross different modalities*.

The paper describes the rationale for why a future version of the TEI guidelines should include models for CMC. It gives an outline of requirements which a framework for the representation of CMC should meet in order to allow corpus providers and researchers to make full use of the abovementioned potentials. It presents an overview of challenges and general issues in designing such a representation framework and thus pre-structures the presentation of models and practices that will be presented in paper 2 as well as the following discussion.

Panel slot #2: *Paper:*

Modeling social media and CMC genres in TEI: Models and practices from French and German corpus projects (40 minutes)

Michael Beißwenger[1], Thierry Chanier[2], Eric Ehrhardt[3], Alexander Geyken[4], Axel Herold[4], Marc Kupietz[5], Lothar Lemnitzer[4], Harald Lüngen[5], Céline Poudat[6], Angelika Storrer[4], Andreas Witt[5]

[1] TU Dortmund University (D) [2] Université Blaise Pascal, Clermont-Ferrand (F) [3] University of Mannheim (D) [4] Berlin-Brandenburg Academy of Sciences and the Humanities (D) [5] Institut für deutsche Sprache, Mannheim (D) [6]University of Nice Sophia Antipolis (F)

The second paper discusses how the requirements and challenges outlined in paper 1 have been handled in customized TEI schemas that have been developed for the representation of CMC and social media genres in French and German corpus projects in 2011–2015. The schemas developed in these projects are not independent from each other but relate to each other: Fostered by discussions in the TEI-SIG "computer-mediated communication", in the German DFG network *Empirikom*[2] and in the French corpus network *CoMeRe*, the projects recursively have been building on each other's work with the goal of creating a schema that fits for diverse projects in several languages:

(1) A first TEI schema for CMC (Beißwenger et al. 2012) has been developed as part of the exploratory work for a reference corpus of German CMC as part of the DWDS corpus collection at the BBAW Berlin (*DeRiK*, Beißwenger et al. 2013).

(2) Margaretha & Lüngen (2014) adopted the basic models introduced in (1) and tested their suitability for the annotation of a corpus of Wikipedia talk pages as part of the DEREKO corpus collection at IDS Mannheim.

(3) Building on the results of two meetings of the TEI-CMC-SIG in Rome (2013) and Dortmund (2014) and on requirements from corpora collected in the French *CoMeRe* network, Chanier et al. (2014) developed a TEI schema which significantly expanded the models suggested in (1) for a corpus collection with highly heterogeneous genres (covering a broad range from text.-only to multimodal genres). This schema has been used for the representation of the CoMeRe repository of French CMC corpora (access to corpora via CoMeRe, 2015).

(4) The schema developed for the CoMeRe corpora (3) as well as the experiences from (2) are presently used as the starting point for defining a schema for the use in several German CMC corpus projects: the CLARIN-D curation project *ChatCorpus2CLARIN* in which the *Dortmund Chat Corpus*[3] is being re-modeled in TEI; a corpus of German Usenet postings which is presently being collected for

[2] http://www.empirikom.net

[3] http://chatkorpus.tu-dortmund.de/

integration in DEREKO (Schröck, in prep.); a corpus of German WhatsApp messages that has been collected in 2014/15[4]. The schemas used for the Wikipedia corpus in DEREKO and in the DeRiK project (see above) shall subsequently be adapted to this new schema version.

The schema versions (3) and (4) will be documented in the TEI wiki before the conference.

Panel slot #3: *Discussion:*

Towards a basic schema for the representation of CMC in TEI
(30 minutes)

With respect to the goals outlined in the introduction of this proposal, the panel includes a 30-minute space for discussions instead of a third paper. The discussion shall be introduced by short statements of two invited discussants who bring in the perspective of modeling related genres and text types in TEI and of "experts" in the process of discussing and implementing new features into the TEI guidelines:

- *Peter Stadler*, Carl-Maria-von-Weber-Gesamtausgabe, Detmold
  (member of the TEI Technical Council / TEI-SIG "Correspondence")
- *Thomas Schmidt*, Institut für deutsche Sprache, Mannheim
  (representation of spoken language corpora / transcribed speech)

The statements by the discussants will be followed by a moderated open discussion with the plenary (which may continue at the meeting of the SIG "computer-mediated communication").

A documentation and ODD files of the schemas presented in paper 2 will be made available in the TEI wiki and be announced via the TEI mailing list before the conference in order to allow the discussants and other participants to examine the CMC models in advance.

References

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation of Computer-mediated Communication. Journal of the Text Encoding Initiative (jTEI) 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2013): DeRiK: A German Reference Corpus of Computer-Mediated Communication. In: Literary and Linguistic Computing (LLC). http://www.tinyurl.com/derik-llc

Chanier, Thierry; Poudat, Celine; Sagot, Benoit; Antoniadis, Georges; Wigham, Ciara; Hriba, Linda; Longhi, Julien; Seddah, Djamé (2014): The CoMeRe corpus for

[4] Project „Whats Up, Deutschland" (http://www.whatsup-deutschland.de/), initiated and coordinated by Beat Siebenhaar (University of Leipzig).

French: structuring and annotating heterogeneous CMC genres. In: Beißwenger, Michael; Oostdijk, Nelleke; Storrer, Angelika; van den Heuvel, Henk (Eds.): Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Special Issue, Journal of Language Technology and Computational Linguistics (JLCL 2/2014), 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf

CoMeRe (2015). CoMeRe Repository: Corpora of Computer-Mediated Communication in French. Ortolang : Nancy. http://hdl.handle.net/11403/comere

Margaretha, Eliza; Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: Beißwenger, Michael; Oostdijk, Nelleke; Storrer, Angelika; van den Heuvel, Henk (Eds.): Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Special Issue, Journal of Language Technology and Computational Linguistics (JLCL 2/2014), 59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf

Schröck, Jasmin (in prep.): Erstellung eines deutschsprachigen Usenet-Newsgroup-Korpus und Annotation von Phänomenen internetbasierter Kommunikation. Universität Heidelberg.

[TEI P5] TEI Consortium (eds) (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange. http://www.tei-c.org/Guidelines/P5/ (accessed 22 March 2013).

About the authors

Michael Beißwenger is a researcher and lecturer for German Linguistics at TU Dortmund University and convener of the TEI-SIG “computer-mediated communication”. Together with Angelika Storrer he leads the CLARIN-D curation project “ChatCorpus2CLARIN”. He is a member of the project group of the DeRiK project (“Deutsches Referenzkorpus zur internetbasierten Kommunikation”).
E-Mail: michael.beisswenger@tu-dortmund.de

Thierry Chanier is a professor of linguistics at the University Blaise Pascal, France. He is a member of the board of the national consortium on written corpora (*Corpus-Ecrits*, component of *Huma-Num*, national research infrastructure linked to DARIAH). He coordinates the *CoMeRE* project (repository of French CMC corpora).
E-Mail: thierry.chanier@univ-bpclermont.fr

Eric Ehrhardt is a researcher in the CLARIN-D curation project “ChatCorpus2CLARIN” at the University of Mannheim.
E-Mail: eric.ehrhardt@gmx.de

Alexander Geyken is a researcher at the Berlin-Brandenburg Academy of Sciences (BBAW) where he is head of the project group of the „Digital Dictionary of German Language“ (DWDS), a long-term project of the BBAW. He is a member of the project group of the DeRiK project (“Deutsches Referenzkorpus zur internetbasierten Kommunikation”) and of the CLARIN-D curation project “ChatCorpus2CLARIN”.
E-Mail: geyken@bbaw.de

Axel Herold is a researcher at the Berlin-Brandenburg Academy of Sciences (BBAW). He is a member of the CLARIN-D center at the BBAW and the CLARIN-D curation project “ChatCorpus2CLARIN”. For the CLARIN ERIC, he serves as national metadata and concept registry coordinator.

E-Mail: herold@bbaw.de

Marc Kupietz is head of the section „Corpus Linguistics“ at the Institut für Deutsche Sprache (IDS), Mannheim, and responsible for the project „Deutsches Referenzkorpus“ (DEREKO). Together with Andreas Witt he leads the project „Korpusanalyseplattform der nächsten Generation“ (KorAP). He is a member of the CLARIN-D curation project “ChatCorpus2CLARIN”.
E-Mail: kupietz@ids-mannheim.de

Lothar Lemnitzer is a lexicographer and researcher in the project group of the „Digital Dictionary of German Language“ (DWDS) at the Berlin-Brandenburg Academy of Sciences (BBAW). He is a member of the project group of the DeRiK project (“Deutsches Referenzkorpus zur internetbasierten Kommunikation”) and of the CLARIN-D curation project “ChatCorpus2CLARIN”.
E-Mail: lemnitzer@bbaw.de

Harald Lüngen is a researcher in the corpus linguistics’ section and a member of the project group of the „Deutsches Referenzkorpus“ (DEREKO) at the Institut für Deutsche Sprache (IDS), Mannheim. He is responsible for the Wikipedia corpus in DEREKO and is a member of the CLARIN-D curation project “ChatCorpus2CLARIN”.
E-Mail: luengen@ids-mannheim.de

Céline Poudat is a researcher and lecturer in corpus linguistics at the University of Nice Sophia Antipolis. She is a member of the board of the national consortium on written corpora (*Corpus-Ecrits*) and participates in the coordination of the *CoMeRe* project.
E-Mail: cpoudat@gmail.com

Angelika Storrer is a full professor and head of the department of German Linguistics at the University of Mannheim. Her current research interests cover corpus linguistics, e-lexicography, and linguistic aspects of computer-mediated communication and social media. She is a member of the board of directors of the German Society of Computational Linguistics GSCL, and a member of the Berlin-Brandenburg academy of Sciences and Humanities (BBAW).
E-Mail: astorrer@mail.uni-mannheim.de

Andreas Witt is head of the section „Research Infrastructures“ at the Institut für Deutsche Sprache (IDS), Mannheim, and honorary professor for Digital Humanities at the University of Heidelberg. He is the convener of the ISO working group „Linguistic Annotation“ (ISO/TC 37/SC 4/WG 6) and of the TEI-SIG „TEI for Linguists“. He is a member of the CLARIN-D curation project “ChatCorpus2CLARIN”.
E-Mail: witt@ids-mannheim.de