Published in: Resch, Claudia/Hannesschläger, Vanessa/Wissik, Tanja (eds.): TEI Conference and Members' Meeting 2016. Book of Abstracts. Vienna: Austrian Centre for Digital Humanities, Austrian Academy of Sciences, 2016. p. 39-41.

Michael Beißwenger, Eric Ehrhardt, Axel Herold, Harald Lüngen, Angelika Storrer

Converting and Representing Social Media Corpora into TEI: Schema and best practices from CLARIN-D

The paper presents results from a curation project within CLARIN-D, in which an existing 1MWord corpus of German chat communication[1] has been integrated into the DEREKO[2] and DWDS[3] corpus infrastructures of the CLARIN-D centres at the Institute for the German Language (IDS, Mannheim) and at the Berlin-Brandenburg Academy of Sciences (BBAW, Berlin).[4] The focus is on the solutions developed for converting and representing the corpus in a TEI format.

The corpus, which has been collected and built in 2002-2008, has originally been annotated using a home-grown XML format that describes the main structural features of chat log files and user postings as well as selected linguistic phenomena of computer-mediated communication (CMC). In order to ensure the sustainability of the resource and its interoperability with the corpus collections already available in CLARIN-D, one important subtask of the project was to define a schema and workflow for remodeling the resource in TEI. Since TEI P5 in its current version doesn't include any models for the representation of CMC

[1] The Dortmund Chat Corpus. http://www.chatkorpus.tu-dortmund.de [2016-07-31].

[2] http://www1.ids-mannheim.de/direktion/kl/projekte/korpora.html [2016-07-31].

[3] http://dwds.de/ressourcen/korpora/ [2016-07-31].

[4] Project description. http://de.clarin.eu/en/curation-project-1-3-german-philology [2016-07-31].

and social media genres, the project adopted and extended the modeling suggestions which have been defined and discussed in previous work of the TEI-SIG "computer-mediated communication (CMC)"[5] and defined a workflow for the automatic, lossless conversion of the source into the target schema.

The target schema[6] has been tested not only with data from the chat corpus, but also with data from a range of other types of CMC and social media genres (whatsapp interactions, wikipedia talk pages, tweets, usenet posts) in order to provide a useful solution for the encoding of other corpora of that type as well. The schema and conversion workflow will be used for the integration of more CMC and social media corpora into the CLARIN-D infrastructures in the near future.

References

Beißwenger, Michael; Ermakova, Maria; Geyken, Alexander; Lemnitzer, Lothar; Storrer, Angelika (2012): A TEI Schema for the Representation

[5] Michael Beißwenger; Maria Ermakova; Alexander Geyken; Lothar Lemnit er; Angelika Storrer (2012): A TEI Schema for the Representation of Computermediated Communication. In Journal of the Text Encoding Initiative 3. http://jtei.revues.org/476 [2016-07-31]; Thierry Chanier; Celine Poudat; Benoit Sagot; Georges Antoniadis; Ciara Wigham; Linda Hriba; Julien Longhi; Djamé Seddah (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In Journal of language Technology and Computational Linguistics (JLCL) 29/2. 1-30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf [2016-07-31]; Eliza Margaretha; Harald Lüngen (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In Journal of language Technology and Computational Linguistics (JLCL) 29/2. 59-82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf [2016-07-31].

[6] The ODD and RNG file for the schema are available in the TEI wiki at http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema [2016-07-31].

of Computermediated Communication. In Journal of the Text Encoding Initiative 3. http://jtei.revues.org/476 [2016-07-31].

Chanier, Thierry; Poudat, Celine; Sagot, Benoit; Antoniadis, Georges; Wigham, Ciara; Hriba, Linda; Longhi, Julien; Seddah, Djamé (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In Journal of Language Technology and Computational Linguistics (JLCL) 29/2. 1-30. http://www.jlcl. org/2014_Heft2/1Chanier-et-al.pdf [2016-07-31].

Margaretha, Eliza; Lüngen, Harald (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In Journal of Language Technology and Computational Linguistics (JLCL) 29/2. 59-82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf [2016-07-31].