(Best) Practices for Annotating and Representing CMC and Social Media Corpora in CLARIN-D

Michael Beißwenger*, Eric Ehrhard†, Axel Herold¥, Harald Lüngen‡, Angelika Storrer†

* Department of German Studies, University of Duisburg-Essen, Berliner Platz 6–8, D-45127 Essen
† Department of German Linguistics, University of Mannheim, Schloss, Ehrenhof West, D-68131 Mannheim
¥ Berlin-Brandenburg Academy of Sciences and Humanities, Jägerstraße 22/23, D-10117 Berlin
‡ Institute for the German Language, R5, 6–13, D-68161 Mannheim

E-mail: michael.beisswenger@uni-due.de, eric.ehrhardt@gmx.de, herold@bbaw.de, luengen@ids-mannheim.de, astorrer@mail.uni-mannheim.de

Abstract

The paper reports the results of the curation project *ChatCorpus2CLARIN*. The goal of the project was to develop a workflow and resources for the integration of an existing chat corpus into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences (http://clarin-d.de). The paper presents an overview of the resources and practices developed in the project, describes the added value of the resource after its integration and discusses, as an outlook, to what extent these practices can be considered *best practices* which may be useful for the annotation and representation of other CMC and social media corpora.

Keywords: CMC corpora, TEI encoding, tagging, corpus infrastructures, legal issues, CLARIN

1. Introduction

This paper reports the results of the curation project *ChatCorpus2CLARIN*. The goal of the project was to develop a workflow and resources for the integration of an existing chat corpus (the *Dortmund Chat Corpus*, Beißwenger 2013) into the CLARIN-D research infrastructure for language resources and tools in the Humanities and the Social Sciences[1] as part of the European *Common Language Resources and Technology Infrastructure*[2]. The paper presents an overview of the resources and practices developed in the project, describes the added value of the resource after its integration and discusses, as an outlook, to what extent these practices can already be considered as *best practices* which may be useful for the annotation and representation of other CMC and social media corpora.

2. Goals of the Project

The goal of the project was twofold: On the one hand, (1) the project aimed to integrate an existing chat corpus into the CLARIN-D corpus infrastructures at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and at the Institute for the German Language (IDS), Mannheim. This included, as subtasks, (1a) the development of a schema and conversion routine for the transformation of the XML markup and metadata in the original resource into a TEI format, (1b) the addition of a new annotation layer with part-of-speech and lemma information, (1c) a re-anonymization of the corpus data according to the recommendations given in a legal opinion. On the other hand, (2) the solutions developed to achieve goal (1) should be designed as general (and not idiosyncratic) approaches to the challenge of annotating and representing corpora of computer-mediated communication (CMC) and social media according to existing standards in the Digital Humanities / CLARIN context. The main result of goal (1) is, thus, the integrated chat corpus whereas the results of goal (2) are documented resources and practices that may be reused by other projects which aim at integrating CMC and social media resources into CLARIN.

3. The Corpus

The *Dortmund Chat Corpus* (Beißwenger, 2013) has been collected at TU Dortmund University as a resource for researching the peculiarities and linguistic variation in written CMC. The corpus comprises 478 chat documents (*logfiles*) containing 140240 user postings or 1M words of German chat discourse from heterogeneous sources representing the use of chats in a wide range of application contexts (social chats, advisory chats, chats in the context of learning and teaching, moderated chats in the media context). The corpus has been annotated using a homegrown XML format ('ChatXML') that describes (1) the basic structure and properties of chat logfiles and postings, (2) selected "netspeak" phenomena such as emoticons, interaction words, addressing terms, nicknames and acronyms, (3) selected metadata about the chat platforms and chat users. Since 2005, a large subset of the corpus has been available as a ChatXML resource for download and offline querying, and as an HTML version for online browsing.[3]

4. Overview of Workflow and Resources

The Dortmund Chat Corpus served as a use case to demonstrate how an integration of CMC and social media resources could be accomplished in a way that the target resource (1) conforms to established stan-

[1] http://clarin-d.de
[2] https://www.clarin.eu
[3] http://www.chatkorpus.tu-dortmund.de

dards for the representation and linguistic annotation of corpora in the Digital Humanities context and (2) can be used for comparative analyses with other types of corpus resources in CLARIN-D (text and speech corpora). A visualization of the workflow and practices developed in the project is given in Fig. 1; the steps and resources of the pipeline are described in the following subsections.

4.1 Experimental CMC Corpus with Data Samples from Heterogeneous Sources

For developing and testing the solutions for goals (1a), (1b) and (1c) (cf. Sect. 2) not only with chat data, we compiled a small experimental corpus of 38382 tokens with data also from other CMC and social media genres. The corpus included (1) two logfiles from different subcorpora of the chat corpus (12526 tokens), (2) 94 news messages from the Usenet corpus in DEREKO (Schröck & Lüngen, 2015) (9108 tokens), (3) excerpts from two Wikipedia talk pages (907 tokens), (4) donated tweets from two different twitter accounts (1412 tokens) and (5) 1907 posts from two different whatsapp conversations collected in the project "What's up, Deutschland?"[4] (14429 tokens).

4.2 The NLP Toolchain Developed in the BMBF Project www.schreibgebrauch.de

Part-of-speech (PoS) tagging was done in two stages: (1) an automatic tagging process and (2) a manual post-editing phase. Automatic tagging (including tokenization, PoS tagging and lemmatization) was done at Saarland University applying an NLP toolchain that was developed in the BMBF project *Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen*[5] (Horbach et al., 2014). The toolchain had originally been trained on annotating chat and forum data with a tag set derived from Bartz et al. (2014).

4.3 The 'STTS 2.0' Part-of-Speech Tagset and Guidelines from the EmpiriST2015 Shared Task Project

As target standard for the PoS layer, we used the STTS-IBK tag set ('STTS 2.0') developed in the GSCL shared task on automatic linguistic annotation of CMC and social media (EmpiriST2015)[6]. 'STTS 2.0' is an advanced version of the tag set suggested in Bartz et al. (2014) and builds on the categories of the "Stuttgart-Tübingen Tagset" (*STTS*, Schiller et al., 1999) which is a well-acknowledged defacto standard for PoS tagging of German written corpora. In its canonical version, STTS does not include any tags for CMC and social media genres. 'STTS 2.0' therefore introduces two types of new tags: (1) tags for phenomena which are specific for CMC and social media discourse, (2) tags for phenomena which are typical of spontaneous spoken language in colloquial registers and which can also be found in corpora of transcribed speech (e.g., in the FOLK corpus of spoken language at the IDS which uses an STTS extension which is compatible with 'STTS 2.0', Westpfahl, 2014). The resulting tag set is still downwardly compatible with STTS (1999) and therefore allows for interoperability with other corpora

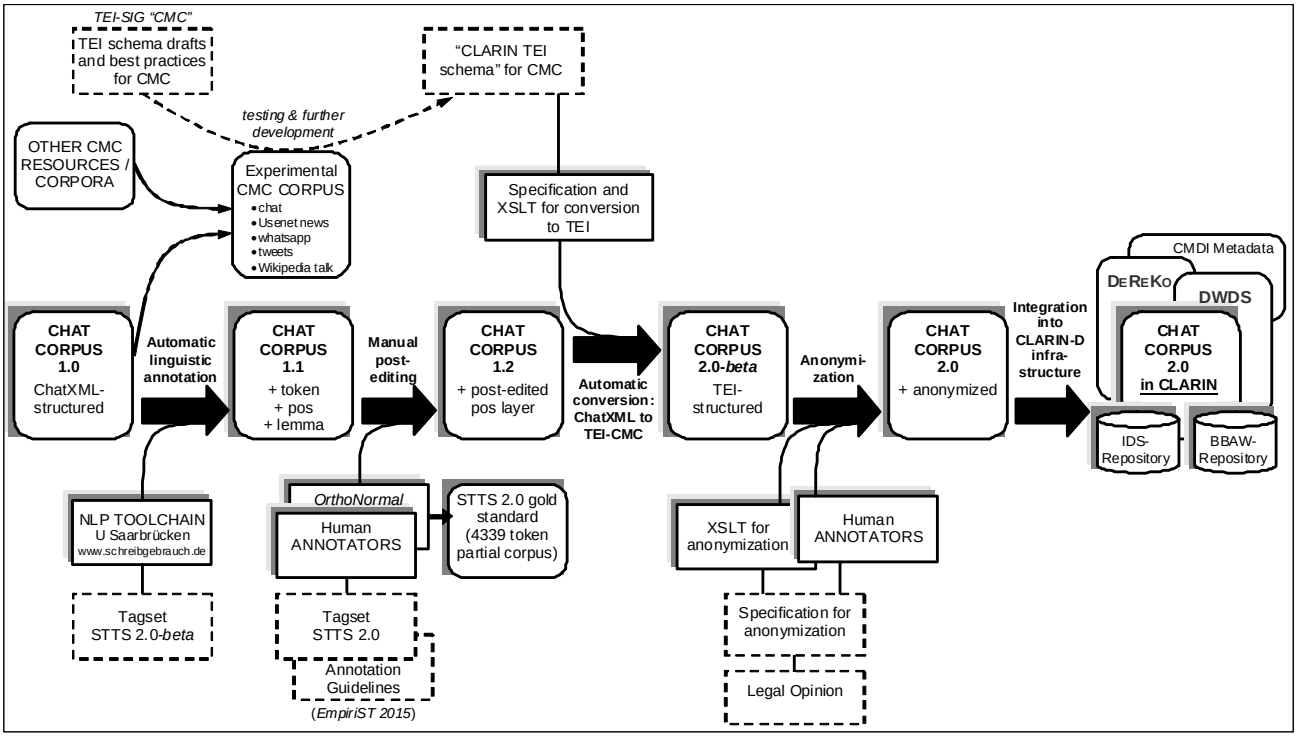

Figure 1: Workflow and resources.

[4] http://www.whatsup-deutschland.de

[5] http://www.schreibgebrauch.de

[6] http://sites.google.com/site/empirist2015/

that have been tagged with STTS. In the EmpiriST2015, several existing NLP systems have been trained on assigning the 'STTS 2.0' extensions to tokens of CMC and social media discourse (Beißwenger et al., 2016). The tag set is described in an annotation guideline (Beißwenger et al. 2015) and had previously been tested with data from several CMC genres. In the curation project, these guidelines have been used for manual post-editing the results of the automatic tagging process described in Sect. 4.2. In the post-editing process which was done using an adapted version of the tool *OrthoNormal* from the *FOLKER* tool suite (Schmidt, 2012), the whole corpus has been made compatible with the 'STTS 2.0' tag set. In addition, for a partial corpus of 4339 tokens all tags assigned in the automatic process have been post-edited independently by two human annotators who had been trained with the guidelines (agreement according to Cohens Kappa: $\kappa = 0.92$). Differing cases were decided by the project heads. The 4339 partial corpus with manually checked PoS annotation can be considered as an additional resource from the project which can be used for further retraining of tagging systems with 'STTS 2.0'.

4.4 The 'CLARIN TEI Schema for CMC' and the XSLT for Conversion

The resource was converted into a TEI representation format which builds on (1) the official TEI-P5 framework for electronic text encoding and interchange and (2) two versions of a customization of TEI-P5 for CMC genres created in the context of the TEI special interest group "computer-mediated communication" (CMC-SIG) and described in Beißwenger et al. (2012) and Chanier et al. (2014). Starting from a close evaluation of the most recent version of the customization Chanier et al. (2014), we developed the models and best practices from the TEI CMC-SIG further taking into consideration the genres available in our experimental corpus. The resulting new TEI schema draft – the 'CLARIN TEI schema for CMC' – has been made available for further use and comments in the TEI wiki[7]. The conversion of the ChatXML format into the target TEI format was done using an XSLT stylesheet.

4.5 Representation of Metadata in TEI

In contrast to the customizations needed for the markup of the primary discourse data, we did not modify the existing TEI metadata model. All metadata provided in the original version of the corpus (which was partially given as part of the ChatXML structure, partially as textual descriptions provided in the corpus-external documentation of the corpus data) could be re-modelled using their TEI equivalents within the teiHeader. Special attention was paid to the modeling of a text classification scheme which is associated with the corpus documents by means of the TEI's generic textClass/catRef mechanism. This model can be easily extended to a broader range of text and/or discourse properties to account for more detailed classifications, such as the one proposed by Herring (2007) – work that hasn't been done within the project but which is a goal for a future extension of the schema.

4.6 Legal Opinion on Republishing the Resource in CLARIN-D – and Consequences (Anonymization)

Prior to the integration of the curated resource in CLARIN infrastructures, we sought a legal opinion to get a better picture of the legal conditions for republishing the material as a whole or in parts. The legal opinion which was provided by *iRights.Law*/John H. Weitzmann (iRights.Law, 2016) carefully checked possible restrictions arising from individual property rights, copyrights and other legal statutes. One result was that the possibility to identify individuals from their utterances (with the exception of public figures) needed to be circumvented by means of an anonymization of names, nicknames, host names and IP addresses, geographical names (e. g. address data) etc. In addition, it turned out that some (minor) parts of the resource must not be made available to the public at all, notably those parts where personality rights of participants are strongly affected. This applies to a subcorpus obtained from chat-based psycho-social counseling (a subcorpus which hadn't been made available to the public even in the original version of the corpus). For this subcorpus, due to the personal context represented in the discourse, anonymization alone is unlikely to prevent the identification of individuals. Consequently, these resources (8 logfiles containing 88227 tokens) were removed from the final corpus.

The legal opinion saw no indication of concerns regarding copyright (German "Urheberrecht", specifically) as it acknowledges that the collected logfiles as well as the individual user posts in the overwhelming majority of cases do not represent works of art. Protectable under EU (and German) law however, is the work committed in the course of collection, curation and transformation of the data into the format of the intended linguistic database. Therefore and in accordance with our goal to provide the resource as openly as possible, we followed the lawyers' suggestion to provide the resource with a CreativeCommons licence (CC BY 4.0) which allows for the protection of database creator rights.

The task of anonymization could not be done completely automatically: In a first step, names that had already been annotated in the original resource could be replaced by categorized placeholders automatically. Likewise, the metadata section and the filenames were anonymized, including names and properties of participants, and the names of chat platforms. What had to be done manually was to replace all those occurrences of names that had not been annotated in the source, or that could not be matched to entries in the participant

[7] http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema

list automatically (e.g., because chatters were addressing each other using nicknames of nicknames or referring to people who were not participating in the chat themselves). This was a very time-consuming process which at the current state could only be done for the 4339 token gold standard subset of the corpus. The anonymization of the rest of the corpus is part of a follow-up work package to be finished at the end of 2016.

5. Availability

All work packages described in Sect. 4.1–4.5 have been finished. Until October 2016, a first release of the resource will present a preview in form of the 4339 token gold standard. It is planned to make the full resource available in a 2nd release in early 2017.

The corpus will be ingested into the CLARIN repositories at the IDS[8]and the BBAW[9]. At IDS, the resource will become part of the German Reference Corpus archive DeReKo and as such will be integrated in the corpus query platform COSMAS II[10]. At BBAW, the corpus will be integrated in the corpus query platform DWDS[11]. In addition, the corpus will be made accessible through CLARIN's federated content search, e.g. for NLP toolchains such as WebLicht[12].

6. Features of the Integrated Resource

Compared with the original version of the resource, the CLARIN-integrated version ('Chat Corpus 2.0', cf. Fig. 1) will allow for advanced queries using the additional linguistic annotations (sentences, tokens, PoS, lemmas). Due to the remodeling of the resource in TEI and the compatibilty of the PoS annotations with STTS the corpus will be interoperable with other TEI-/STTS-annotated language resources. The integration into the CLARIN-D corpus infrastructures at BBAW and IDS will facilitate the comparative analysis of the chat corpus with the BBAW and IDS text and speech corpora. These features will not only increase the value of the resource for language-centered CMC research and variational linguistics but also the possibilities to use it in language teaching and higher education.

7. Outlook

According to goal (2) (cf. Sect. 2), the resources and practices developed in the project were meant to function as general approaches to open issues in representing and annotating CMC and social media data which should have the potential to be useful also for other projects in the field. To assess empirically whether the current versions of the resources (the TEI schema, the 'STTS 2.0' tagset and annotation guidelines) already have this potential, it is necessary to adopt and test these resources in other CMC corpus projects. In our own work, we tested them not only with chat data but also with a selection of data from other genres (experimental corpus, cf. Sect. 4.1). We're optimistic that the availability of the resources will facilitate corpus annotation for colleagues who are building similar corpora and who are aiming to represent them on the basis of existing standards same as we did when adopting the encoding framework of the TEI and the STTS tagset for German for our purpose. We're aware of the fact that no existing schema – not even an established standard as TEI-P5 – can usually be adopted for a new project to 100%; instead, each project typically needs their own customizations and extensions when adopting an existing solution. Nevertheless, customizing and extending a given solution is usually much easier than having to start to design a solution from scratch. Especially the TEI schema for CMC is open for further changes according to experiences and results from other projects. It will be the basis for further discussions in the TEI-SIG "computer-mediated communication" which is open for the participation to everybody who is interested to bring in their own experiences and suggestions.

In our own work, we are planning to adopt the resources and practices from the project for the integration of further CMC and social media resources into the CLARIN-D corpus infrastructures at the IDS and the BBAW (starting as of autumn 2016). The TEI schema, in addition, is currently being used and tested also in projects in which none of the authors of this paper is involved – e.g., in a weblog corpus project at the University of Gießen, Germany ('Discourse-structured Blog Corpus for German', Karlova-Bourbonus et al., 2016) and for the annotation of an English Q&A corpus at the University of California, Davis, USA (Rachael Duke, Raul Aranovich).

The 'STTS 2.0' tagset for PoS tagging CMC and social media data has been used for the EmpiriST2015 shared task in which several NLP systems have been adapted for the automatic annotation of German CMC. These systems will allow corpus projects to achieve better results in tagging their data than with standard NLP tools which have typically been trained only on 'standard' genres (newspaper corpora etc.).

The results and recommendations of the legal opinion will be a useful point of reference for further inquiries into the (still difficult) legal conditions of collecting and republishing discourse from CMC and social media sources as parts of linguistic research infrastructures.

8. References

Bartz, T., Beißwenger, M., Storrer, A. (2014). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguis-*

[8] https://repos.ids-mannheim.de/
[9] http://clarin.bbaw.de/en/repo/
[10] http://cosmas2.ids-mannheim.de/
[11] http://www.dwds.de/
[12] https://weblicht.sfs.uni-tuebingen.de/weblicht/

tics 28 (1), pp. 157–198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik* 41 (1), pp. 161–164. http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

Beißwenger, M., Bartsch, S., Evert, S., Würzner, K.-M. (2016). EmpiriST 2015: A Shared Task on Automatic Linguistic Annotation of Computer-Mediated Communication, Social Media and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26), 44–56. http://aclweb.org/anthology/W16-26

Beißwenger, M., Bartz, T., Storrer, A., Westpfahl, S. (2015). *Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation.* Guideline document from the EmpiriST2015 shared task. http://sites.google.com/site/empirist2015/home/annotation-guidelines

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., Storrer, A. (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).

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C., Hriba, L., Longhi, J., Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: *Journal of language Technology and Computational Linguistics (JLCL) 29 (2)*, pp. 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf

Herring, S.C. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet 4 (1)*. http://www.languageatinternet.org/articles/2007/761

Horbach, A., Steffen, D., Thater, S., Pinkal, M. (2014). Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In *Proceedings of KONVENS 2014*, pp. 171–177.

iRights.Law Rechtsanwälte (2016). *Rechtsgutachten zur Integration mehrerer Text-Korpora in die CLARIN-D-Infrastrukturen.* (Legal opinion for the ChatCorpus2CLARIN project, 46 pages)

Karlova-Bourbonus, N., Grumt Suárez, H., Lobin, H. (2016). Compilation and Annotation of the Discourse-structured Blog Corpus for German. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*, University of Ljubljana [this volume].

Schiller, A., Teufel, S., Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).* University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmidt, T. (2012). EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12)*. http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf

Schröck, J., Lüngen, H. (2015). Building and Annotating a Corpus of German-Language Newsgroups. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015)*, pages 17–22. https://sites.google.com/site/nlp4cmc2015/proceedings.

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

Westpfahl, S. (2014). STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*. Association for Computational Linguistics (ACL Anthology W14-49), 1–10. http://www.aclweb.org/anthology/W14-4901