Published in: Dipper, Stefanie/Neubarth, Friedrich/Zinsmeister, Heike (eds.): Proceedings of the 13th Conference on Natural Language Processing (KONVENS). Bochum: Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, 2016. p. 156-164. (Bochumer Linguistische Arbeitsberichte 16)

Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN

Harald Lüngen
Institut für Deutsche Sprache
luengen@ids-mannheim.de

Michael Beißwenger
Universität Duisburg-Essen
michael.beisswenger@uni-due.de

Eric Ehrhardt
Universität Mannheim
eric.ehrhardt@gmx.de

Axel Herold
Berlin-Brandenburgische Akademie der Wissenschaften
herold@bbaw.de

Angelika Storrer
Universität Mannheim
astorrer@mail.uni-mannheim.de

Abstract

We introduce our pipeline to integrate CMC and SM corpora into the CLARIN-D corpus infrastructure. The pipeline was developed by transforming an existing CMC corpus, the Dortmund Chat Corpus, into a resource conforming to current technical and legal standards. We describe how the resource has been prepared and restructured in terms of TEI encoding, linguistic annotations, and anonymisation. The output is a CLARIN-conformant resource integrated in the CLARIN-D research infrastructure.

1 Introduction

Written language in computer-mediated communication (henceforth CMC) and social media (SM) is an important type of non-standard language usage. Although there has been a lot of research on CMC and SM genres in linguistics and social sciences, most of these studies rely on small datasets or corpora that are not publically available. It would be highly desirable to integrate more CMC and SM corpora in corpus collections and to set up common standards for the representation and annotation of these new forms of communication and their structural and linguistic peculiarities.

The project Chatcorpus2CLARIN aimed to explore the prerequisites for integrating CMC und SM corpora into the CLARIN-D corpus infrastructure by transforming an existing CMC corpus, the Dortmund Chat Corpus, into a resource that conforms to current corpus standards. This integration will allow for a systematic corpus-based analysis of CMC and SM discourse as compared with discourse in edited text (as represented in the text corpora at the CLARIN-D centres Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) and Institut für Deutsche Sprache, Mannheim (IDS)) and to spoken conversations (as represented in the spoken language corpora at IDS). The method of transformation developed for this curation project, which is described in this paper, is regarded as a model for the CLARIN curation of CMC corpus resources in general. (Thus, throughout this paper, the term *curation* is used in the concrete sense of "CLARINification".)

The paper is structured as follows: In the following section we provide information on the curated resource (Chat Corpus 1.0) and discuss some legal issues that had to be considered in the context of the curation. In our main Section 3, we describe how this resource has been restructured to conform to current standards for the representation of corpora in the Digital Humanities context. In Section 4, we describe the resulting resource Chat Corpus 2.0 and outline the added values that will be created by integrating this resource into the CLARIN infrastructure.

2 Resource and conditions

2.1 The resource

The Dortmund Chat Corpus (Beißwenger, 2013) has been collected at Dortmund Technical University between 2000 and 2006 as a resource for researching the peculiarities and linguistic variation in written computer-mediated communication. The corpus comprises 478 chat documents (logfiles) with 140,240 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 in ChatXML, for download and offline querying and as an HTML version for online browsing[1]. It has been widely used as a resource for studying and teaching the characteristics of German CMC discourse.

2.2 Legal issues

Prior to the integration of the curated resource in CLARIN infrastructures, we sought a legal opinion to decide on questions regarding the republishability of the material as a whole or in parts, i.e. the provisions needed with respect to questions of copyright and personality rights as well as questions regarding the licensing of the corpus.

The corpus comprises personal communication in both private, educational and public chatrooms. To prevent the public revelation of participants' personal data, the possibility to identify individuals from their utterances (with the exception of public figures) needs to be circumvented as much as possible. This is achieved by means of the anonymisation of names, nicknames, host names and IP addresses, geographical names (e.g. address data) etc. (see Section 3.4 for a technical discussion of the anonymisation performed). In accordance with the legal opinion, some parts of the resources data must not be made available to the public at all, notably those parts where personality rights of participants are strongly affected. This applies to all data obtained from chat-based psycho-sociological counseling services in the original corpus (8 chat logfiles with in sum 88227 tokens). Here, due to the highly personal context represented in the discourse, anonymisation measures alone are unlikely to prevent the identification of individuals. Consequently, these resources 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 discourses and the single user contributions in the overwhelming majority of cases do not represent works of art. Protectable under 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 and provided the resource with a CreativeCommons licence (CC BY 4.0), which allows for the protection of database creator rights.

[1] from http://www.chatkorpus.tu-dortmund.de

3 Method

One goal of the project was to develop a model for the integration of CMC and SM corpora into the CLARIN-D corpus infrastructures at BBAW and IDS. The Dortmund Chat Corpus served as a use case to demonstrate how such an integration could be accomplished in a way that the target resource (1) conforms to established standards 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 visualisation of the workflow developed in the project is shown in Figure 1; the steps and resources of the pipeline are described in the following subsections.

3.1 TEI representation

For many years, the guidelines of the Text Encoding Initiative (TEI) have been the de facto standard framework for text (and text structure) encoding in the Digital Humanities. Consequently, the TEI guidelines serve as a suggested best practice in the CLARIN-D corpus research infrastructure (CLARIN-D AP 5, 2012) for different text types, such as historical and contemporary books, newspapers, and other printed resources. However, when trying to model CMC in TEI, there are two fundamental challenges: Firstly, as argued above, CMC shares characteristics with both text and spoken conversation. On the one hand, CMC constitutes dialogic interaction in which each communicative move creates or changes the context for follow-up moves. On the other hand, written CMC is organised through the exchange of stretches of written text which have been completed before they are transmitted and read. A basic model for the representation of user contributions to written CMC (post, s.b.) should reflect these properties. The second challenge is that a basic schema for CMC should be flexible enough to represent multimodal CMC interactions as well, such as the interactions of teachers and students on an e-learning platform. So far, the official TEI P5 Guidelines do not include features that model these basic characteristics, see also Beißwenger et al. (2012).

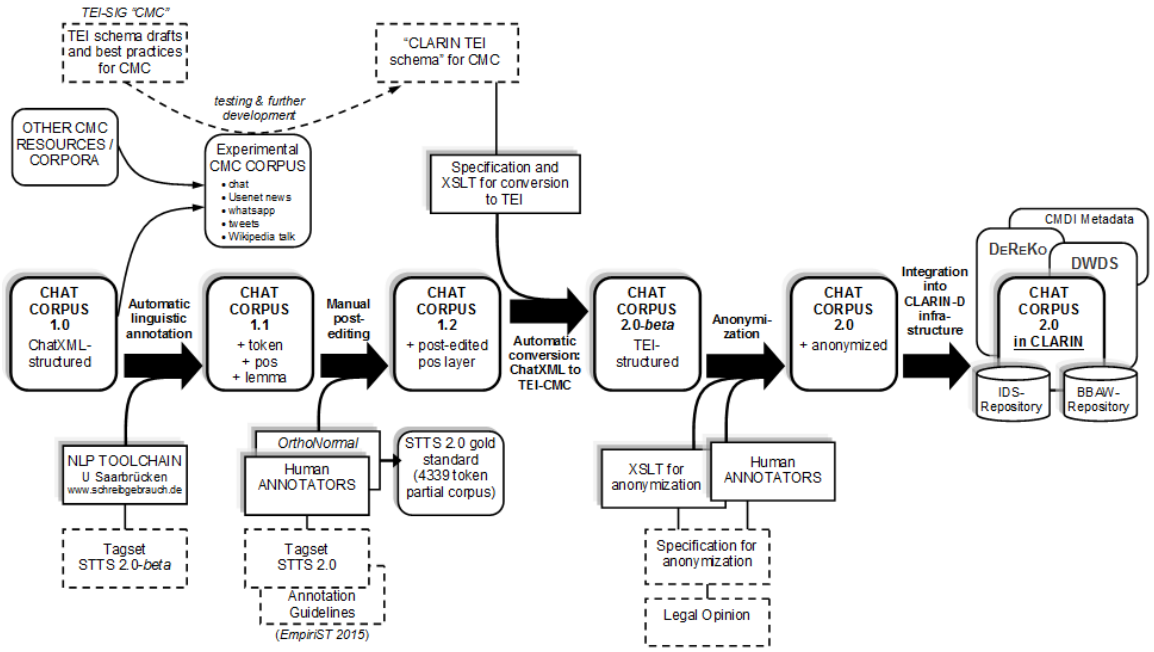

Figure 1: CMC corpus curation pipeline

As a consequence, several corpus initiatives represented in the TEI Special Interest Group Computer-Mediated Communication (TEI CMC SIG) have put forward TEI customisations for different types of CMC and social media genres in the past few years. Two schema drafts resulting from corpus projects in Germany and France have been published by Beißwenger et al. (2012) (DeRiK schema) and Chanier et al. (2014) (CoMeRe schema). The main features of these proposals are the introduction of the elements <post> for written user contributions to CMC interactions, thus combining features of text divisions and spoken utterances (Beißwenger et al., 2012), and <prod> for the representation of non-verbal acts (Chanier et al., 2014). The elements <post>, <prod>, and <u> (the latter marking a spoken utterance in TEI) have been (re-)defined such that they may be combined within one interaction (ibid.).

In the CLARIN-D project, we tested the suitability of the CoMeRe schema for our project by compiling an experimental corpus of German CMC data consisting of chat data (two chat logfiles from the Dortmund Chat corpus), Usenet news (94 news messages from one newsgroup of the Usenet corpus in DeReKo, cf. Schröck & Lüngen (2015)), Wikipedia discussions (five talk pages with 10148 tokens), twitter data (1412 tokens of donated tweets from two different twitter channels), and What'sApp data (1907 messages from the data collected in the project "What's up, Deutschland?"[2]. We then manually annotated the experimental corpus according to the CoMeRe TEI schema and as a result identified a set of CMC features that could not be encoded using it, i.e. for which we had to find new solutions within TEI. Hence, we went for a new, project-specific TEI customisation, dubbed CLARIN CMC-TEI. Our focus was to customise features and to describe best practices for representing the chat data, while the other CMC genres in the experimental corpus were used as supporting or additional evidence.

We decided that for the present project, lexical CMC phenomena such as action words, acronyms, emoticons, and addressing terms are more appropriately annotated on the part-of-speech level, as the tagset STTS 2.0 with corresponding extensions for CMC has recently been introduced cf. Section 3.2), and one tagging system has already been trained for it using CMC data (Horbach et al., 2014), with excellent results on chat data. The POS tags were included in the @type attribute of the <w> elements which mark the tokens. Thus, no TEI customisation would be needed for accommodating these anymore.

The features and solutions of our CLARIN CMC TEI schema are of three types with respect to their relation to the generic TEI P5 guidelines (version 2.9.0):

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

1. Additions of new models for the elements <post>, <prod>, <signatureContent>, and for the two model classes model.floatP.cmc, and model.divPart.cmc.
2. Modifications of existing TEI P5 models so that they fit certain CMC phenomena (e.g. adding @who, @auto; changing <post>, <p>, <s> and <quote>, to include the new model class model.floatP.cmc).
3. "Best practice" solutions using existing TEI P5 models according to specific CMC practice, e.g. use of <w> and <phr> and their attributes for representing word tokens and phrases, respectively, and their POS tags and lemma information.

The first two types are true TEI customisations and have been implemented in our Chat2CLARIN TEI schema. The third type is entirely based on the existing TEI framework and effectively suggests restrictions for the use and application of generic TEI models.

In the following, we explain in more detail one example of each type of solution.

3.1.1 Addition of a new model: <post>

The element <post> models a written contribution to an ongoing CMC interaction which (1) has been composed by its author in its entirety as part of a private activity and (2) has been sent to the server en bloc (Beißwenger et al., 2012).

From the perspective of its addressees/readers, a post is a passage of text that has been composed in advance. Posts occur in a wide range of written CMC genres: as user messages in chats and WhatsApp dialogues, as SMS messages, as tweets in Twitter timelines, as individual comments following a status update on Facebook pages, as posts in forum threads, as contributions on Wikipedia talk pages or in the comments section of a weblog.

The <post> element is provided with five post-specific, optional attributes that serve to model a small set of post metadata: Firstly, @correspAction, which is used to encode the 'sent'/'delivered'/'read' status of a post as in WhatsApp dialogues; the name of the attribute follows the element of the same name in the TEI standard. Secondly, @replyTo indicates to which previous post the current post replies or refers to. The remaining three, @revisedBy, @revisedWhen, and @indentLevel are adapted from the DeRiK-Schema (Beißwenger et al., 2012).

3.1.2 Modification of existing TEI P5 models: The attributes @who and @auto

In the TEI guidelines, the attribute @who indicates the person, or group of people, to whom the element content is ascribed. Besides its application in the <teiHeader>, it is most notably used for references from individual utterances (<u>) to discourse participants (or fictional characters in the case of literary works). As the equivalent to <u> in our schema is <post>, we allow @who for posts in order to indicate post creators. The participants metadata are recorded in a participant list (<particDesc>) within the <profileDesc> section of the <teiHeader> (see Section 3.4 for issues concerning anonymisation) providing each participant with a unique xml:id. The xml:id is then used to establish the reference from posts to participants.

Not all participants in a CMC discourse are necessarily humans. Introduction of automatic chatbots is unproblematic in the adopted framework as they differ from human participants only in their metadata properties but not in their formal behaviour in discourse. However, many mediating systems are able to generate messages or parts of messages on their own, e.g. to indicate that a participant entered or left a chat room or by automatically providing time stamps or signatures in posts. This behaviour of the mediating system is typically triggered by specific actions of the discourse participants. To account for automatically generated parts of messages, an additional attribute @auto with a binary value domain (true, false) was introduced. By combining @who with @auto it becomes possible in principle to model different scenarios of human-machine interaction, including phenomena such as automatic correction of words during typing or the substitution of textual emoticons by their graphical equivalents (@who="HUMAN_PARITICIPANT", @auto="true").

3.1.3 Best practice for CMC: Modelling further aspects of posts in TEI

As can be seen in the example of a post in TEI in Listing 1, certain aspects of a post are modelled using available TEI attributes and elements: The creator of a post is given in the @who attribute, which contains a pointer to the creators entry (<person> element in the participant description in the metadata). Similarly, the posting time (extracted from the timestamp) is given through the reference in the attribute @synch which refers to a point in the

Listing 1: A post element and its annotations

```
<post xml:id="m645" who="#A02" synch="#t058" type="standard" auto="false">
  <note auto="true" who="#A02">for all</note>
  <anchor type="sentence_start"/>
  <ref type="addressingTerm" corresp="#A27">
    <w xml:id="m645.t1" type="ADV" lemma="nun">nun</w>
    <w xml:id="m645.t2" type="VVFIN" lemma="bitten">bitte</w>
    <w xml:id="m645.t3" type="NE" lemma="[_FEMALE-STUDENT-A27_]">[_FEMALE-STUDENT-A27_]</w>
    <w xml:id="m645.t4" type="$." lemma="!">!</w>
  </ref>
  <time> 16:48 </time>
</post>
```

timeline in the metadata section. Note that a timestamp as part of the text is represented in a <time> element, and the string indicating the private/public mode as shown in the original message is annotated by the <note> element. (Similarly, a signature stamp as e.g. used in Wikipedia discussions, would be represented in a <signed> element.) In accordance with the TEI Guidelines, the tokens in our chat corpus (derived from the tokenisation of the Saarland tagging pipeline, cf. Section 3.2.1, are represented by <w> elements. For the inclusion of token-related PoS analyses (including lemma information), there are two basic options offered the TEI by the TEI P5 Guidelines (ch. 17): as inline annotations, i.e. in attributes of <w>, or alternatively, as standoff annotations using the @ana attribute indicating span or feature structure elements elsewhere that contain the analysis. In this project we chose the first method. At <w>, the @lemma attribute contains the lemmatisation info, and the @type attribute contains the POS, see Listing 1. For occurrences of nicknames, chat room names, and addressingTerms, which had been marked up in the original ChatXML, we used the TEI <name> and <ref> elements, with a set of suitable values of their @type attribute ('roomname', 'nickname', 'addressingTerm', and 'url'). In a similar vein, we have introduced many more usage conventions for regular TEI elements and their attributes for the encoding of CMC phenomena.

3.1.4 Best practice for CMC: Metadata

In contrast to the customisations 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 could be 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 texts 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).

However, the TEI guidelines for metadata modeling are currently unable to account for crucial information about properties of the (software) system used to mediate the communication. There are very few means to informally describe the recording equipment used. For CMC systems, a fine-grained formal description of their properties is highly desirable to trace the system's influence on the discourse, especially in large and heterogeneous CMC corpora, possibly comprising multi-modal and/or multi-channel communication. Due to the rapid evolution of CMC systems, it will be difficult for future researchers to take into account relations among the properties and modes of use of a CMC system and properties of the discourse constructed using this system (e.g. communication channels available vs. actually used, automatic transformations of participants' utterances, exact time delays between utterances and their receptions etc.). The discussion of solutions to this problem will be taken up by the TEI special interest group on CMC.

The final CLARIN TEI schema for modeling CMC data according to the solutions developed in our project is publicly available in the form of a documented ODD customisation on the public website of the TEI special interest group on CMC[3].

3.2 Linguistic annotation

Linguistic annotation of the corpus comprised tokenisation, lemmatisation, and part-of-speech (PoS) tagging. While the original ChatXML resource already included annotations for selected CMC phenomena such as emoticons, interaction words, nicknames and addressing terms, one goal of the curation project was to systematically add a layer with PoS annotations in order to extend the

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

possibilities for linguistic queries.

For this purpose, we used the STTS-IBK tag set ('STTS 2.0') from the GSCL shared task on automatic linguistic annotation of CMC and SM genres (EmpiriST2015[4]) which had been defined as a result from discussions in the DFG scientifc network Empirikom[5] and in the context of three workshops dedicated to the adaptation and extension of the canonical version of the Stuttgart-Tübingen-Tagset STTS (Schiller et al., 1999) to the peculiarities of "non-standard" genres (cf. the volume Zinsmeister et al. (2013)). STTS-IBK is a customisation of the canonical STTS version as it 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. The resulting tag set is still backwards compatible with STTS (1999) and therefore allows for interoperability with other corpora that have been tagged with STTS. In addition, the tag set extensions defined in STTS-IBK are compatible with the extensions used at the IDS for the PoS annotation of FOLK, the Mannheim "Research and Teaching Corpus of Spoken German"[6] (Westpfahl, 2014). The tag set is described in an annotation guideline (Beißwenger et al., 2015a) and has been tested with data from several CMC genres in advance. A tabular overview of tags which have been added to the STTS in STTS 2.0 is given in Beißwenger et al. (2015b).

The linguistic preprocessing of the corpus was done in two steps: (1) an automatic step using a toolchain developed at Saarland University (including a basic sentence annotation, tokenisation, PoS and lemma annotation) and (2) a manual step in which the PoS tags resulting from step 1 were post-edited and made compatible with STTS_IBK by two human annotators, cf. Figure 1.

3.2.1 Automatic annotation

The automatic step was carried out by the team of the chair for computational linguistics at Saarland University using the tools for sentence segmentation, tokenisation, PoS tagging and lemmatisation developed in the BMBF project *Schreibgebrauch*[7] and described in (Horbach et al., 2014). These tools were already adapted to the processing for chat and forum data. For the PoS layer they had been trained for assigning the categories of the draft version of STTS 2.0 described in (Bartz et al., 2013) plus some additional categories defined by the developers at Saarland University (tag set STTS 2.0-beta). The result of this automatic tagging process was represented in an extended ChatXML format including token, lemma and PoS information (Tagged ChatXML).

Specific tags in STTS 2.0-beta	Target tags in STTS 2.0
AW	AKW
AWIND	$(
ERRAW	XY
ERRTOK	XY
PROAV	PAV

Table 1: Mapping from STTS 2.0-beta to STTS 2.0

3.2.2 Post-editing of the PoS results

Post-editing included (1) an upgrade of the PoS annotations resulting from step 1 to the STTS 2.0 tag set as described in (Beißwenger et al., 2015b) and, (2) a manual correction of tagging errors in the results from step 1 for a sample of parts of ten chat logfiles, comprising 4,339 tokens altogether.

The manual post-editing of the tagged ChatXML was carried out using the normalisation editor OrthoNormal in FOLKER from the FOLK-Tools Suite (Schmidt, 2012), which was originally developed for the manual normalisation and correction of PoS-tagged spoken language transcripts in the IDS FOLK corpus. For this purpose, Thomas Schmidt (IDS) provided an import and export interface for PoS-tagged ChatXML as part of FOLKER (version 1.2).

In work package (1), the upgrade of the tags used by the Saarland toolchain to STTS 2.0, we mapped specific tags from the Saarland tag set to tags in our target tag set (Table 1). On this basis, all occurrences of the tags in the left column were replaced by the tags in the right column.

Work package (2), the manual correction of tagging errors in the results from step 1, was done independently by two annotators who had been trained on assigning the STTS 2.0 categories beforehand. Based on the EmpiriST2015 guidelines for PoS tagging CMC (Beißwenger et al., 2015a), both annotators checked the PoS tag for each token in a sample comprising approximately 1,000 tokens of data from each of the four top-level text classes of the corpus (social chat, advisory chat, chat in

[4]https://sites.google.com/site/empirist2015/ and cf. Beißwenger et al. (2016)

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

[6]http://agd.ids-mannheim.de/folk.shtml

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

the context of learning, chats in the media context, N=4,339 tokens in ten different logfiles). The tagging results of the two annotators were used for calculating Cohens Kappa ($\kappa = 0.92$). The cases were the annotators had assigned different PoS tags (N=347) were extracted from the ChatXML and presented to the project leaders with a context size of one user posting per token. The project leaders decided the differing cases; on the basis of these decisions, we created the final version of the PoS-tagged ChatXML sample.

An evaluation of the 347 cases in which the tags of the two annotators differed showed that 25,9% of all cases (N=90) could easily be solved with additional restrictions for the use of tags from the canonical STTS (especially of tags for punctuation); the lions share of the remaining cases concerns the distinction between adverbs, modal and gradation particles. Based on these results, further specifications about assigning the STTS 2.0 categories for modal and gradation particles were added to the annotation guidelines.

3.3 ChatXML to TEI Conversion

We implemented a "ChatXML2TEI" XSLT stylesheet to convert the chat documents in Chat Corpus 1.2 (cf. Figure 1), including all metadata and image references, to the CLARIN CMC-TEI format as described in Section 3.1. We also implemented a wrapper script to generate a containing <teiCorpus> element with an appropriate <teiHeader>, combining all the individual chat documents in one large TEI corpus file. The result is the Chat Corpus 2.0 beta, TEI-structured, as indicated in Figure 1.

For quality assurance, we generated a log file of the conversion process, logging e.g. image references, nicknames not matched in the participant list, unusual timestamp formats, unusual element configurations, and the like, and checked it carefully, modifying the XSLT if necessary. We also performed a "primary data diff", i.e. we checked that the raw text contained in the ChatXML files was identical to the raw text of the resulting TEI files, to ensure that the conversion was complete in every case. The single TEI chat files and the combined TEI large corpus file were successfuly validated against the CLARIN CMC-TEI RNG schema using the jing validator[8].

[8] http://www.thaiopensource.com/relaxng/jing.html

NE category	Meaning
PER	Person
ORG	Organisation
LOC	Geographic location
GPE	Geopolitical entity
OTH	Other

Table 2: NE categories according to Telljohann et al. (2004)

3.4 Anonymisation

The obtained legal opinion (cf. Section 2.2) confirmed what is generally known about linguistic data and personality rights: Elements of the data that can be connected to a person (either a chat participant or mentions of a chat-external person) by any means likely reasonably to be used must be anonymised before the data can be published. In practice, this means that linguistic units such as names of persons, places, organisations, but also referring expressions such as URLs or email addresses, including parts that occur only in the metadata such as chatroom names or platform names, need to be obscured. Even indirect sensitive references, such as mentions of the rare hobby of a person, should be anonymised. However, names of politicians and celebrities such as "Sabine Christiansen" (the name of a political talk show host) need not be removed. In order that a corpus can still be reasonably used by linguists after anonymisation, it is recommended that such references are not simply removed but *categorised*, i.e. replaced with a placeholder string expressing the category of the element that has been replaced or even, when more effort can be invested, *pseudonymised,* i.e. "replacing a reference with a variant of the same type", cf. (Medlock, 2006). In the present project, we realised anonymisation as categorisation. Since most of the elements to be anonymised in the chat corpus are names, we used the named entity class set that was used in the Tüba-D/Z treebank (Telljohann et al., 2004), see Table 2.

Since the five categories of this set are rather broad, and in some cases the annotation of the original Chat Corpus 1.0 contained more specific information, we extended the set by the categories NICKNAME (subcategory of PER), and ROOMNAME. The majority of the chat nicknames mentioned is connected via @who or @corresp to the list of creators given in the <particDesc> of the TEI header, so wherever possible we used this entry to derive a more meaningful replacement string,

consisting of a.) the info in the @sex attribute of the participant (i.e. the corresponding <person> in participant description in the TEI header), if available; b.) the info in the @role attribute of the participant, if available, or, if unavailable, the string 'PARTICPANT'; c.) the @xml:id of the participant. The fancy replacement strings such as "FEMALE-TEACHER-A08" were used as replacements in the primary textual data, and in the @lemma and @normal (normalised form) attributes of the <w> elements. A result of this anonymisation procedure can also be seen in Listing 1.

Apart from the content of <name>, we use the the NE replacement categories also for the content of <ref type=addressingTerm>. We have defined further replacement strings for other types of references, e.g. 'WWWURL' and 'EMAIL' for mentions of URLs, or email addresses, respectively.

Anomymisation (i.e. replacing occurrences of names and similar references by the replacement strings described above) was performed in two steps (see Fig. 1):

1. **Automatic** anonymisation using an XSLT stylesheet that operated on the names that had already been annotated and in most cases linked to the creator list of the original resource (using the TEI elements and attributes <name>, <ref>, @who, @corresp, and the <person>s in the header's <particDesc>).
2. **Manual** anonymisation of the remaining occurrences of names that had not been annotated in the source, or that could not be matched in the participant list by the automatic procedure. – However, note that this time-consuming process has only been completed for the tokenised, normalised, and PoS-tagged subset of ten logfiles described in Section 3.2 so far.

4 Result: CLARIN-conformant Resource

The resulting resource is dubbed Dortmund Chat Corpus 2.0, and it contains 470 chat logfiles, containing 131,033 posts, containing 1,005,166 tokens altogether. The file (pretty printed XML) has a size of 100MB. The Dortmund Chat Corpus 2.0 will be ingested in TEI format into the CLARIN repositories at the IDS[9] and the BBAW [10]. At IDS, the chat corpus will become a corpus within the German Reference Corpus archive DeReKo and as such will be integrated in the corpus query platform COSMAS II[11], at BBAW, the corpus will be integrated in the corpus query platform DWDS (Digital Dictionary of the German Language[12]) as of autumn 2016. In addition, access will be provided to it through CLARINs federated content search, e.g. for NLP toolchains such as WebLicht.[13] However, the resource will be fully accessible and downloadable for academic use only when it is completely anonymised. Its complete anonymisation is currently undertaken as a separate effort.

5 Conclusion and prospects

Compared with the previous version of the resource, the Chat Corpus 1.0, the CLARIN-integrated version Chat Corpus 2.0 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 in 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. Last but not least, the schemas, guidelines and best practices developed in the project which are all documented online will be useful resources for the curation of other CMC and SM corpora and their integration in the CLARIN infrastructure. The produced gold standard with PoS-tagged chat data may be used as an additional resource for the further adaptation of NLP tools to the peculiarities of CMC and SM data and corpora. It is planned to apply the pipeline described in this paper (Figure 1) for the remodeling, preprocessing and integration of further CMC and SM corpora in CLARIN in the near future.

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

References

Thomas Bartz, Michael Beißwenger, and Angelika Storrer. 2013. Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. *Journal for Language Technology and Computational Linguistics. Special edition "'Das STTS-Tagset für Wortartentagging – Stand und Perspektiven"', edited by Heike Zinsmeister, Ulrich Heid & Kathrin Beck*, 28(1):157–198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf.

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

Michael Beißwenger, Thomas Bartz, Angelika Storrer, and Swantje Westpfahl, 2015a. *Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline-Dokument aus dem Projekt "'GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication / SocialMedia"' (EmpiriST2015)*. https://sites.google.com/site/empirist2015/home/annotation-guidelines.

Michael Beißwenger, Eric Ehrhardt, Andrea Horbach, Harald Lüngen, Diana Steffen, and Angelika Storrer. 2015b. Adding Value to CMC Corpora: CLARINification and Part-of-Speech Annotation of the Dortmund Chat Corpus. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015)*, pages 12–16, Essen. https://sites.google.com/site/nlp4cmc2015/program.

Michael Beißwenger, Sabine Bartsch, Stefan Evert, and Kay-Michael Würzner. 2016. EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, volume W16-26 of *ACL Anthology*, pages 44–56. Association for Computerational Linguistics, Stroudsberg.

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

Thierry Chanier, Céline Poudat, Benoit Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi, and Djamé Seddah. 2014. The CoMeRe corpus for French: Structuring and annotating heterogeneous CMC genres. *Journal of language Technology and Computational Linguistics*, 29(2):1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf.

CLARIN-D AP 5. 2012. CLARIN-D User Guide. http://de.clarin.eu/de/hilfe/benutzerhandbuch.

Susan C Herring. 2007. A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet*, 4(1):1–37. http://www.languageatinternet.org/articles/2007/761.

Andrea Horbach, Diana Steffen, Stefan Thater, and Manfred Pinkal. 2014. Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. In *Proceedings of KONVENS 2014*, pages 171–177.

Ben Medlock. 2006. An introduction to nlp-based textual anonymisation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1051–1056.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technical report, Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

Thomas Schmidt. 2012. EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language. In *Proceedings of the Eight conference on International Language Resources and Evaluation (LREC12)*, pages 236–240. European Language Resources and Evaluation (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Jasmin Schröck and Harald Lüngen. 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, Essen. https://sites.google.com/site/nlp4cmc2015/program.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004*, pages 2229–2235.

Swantje Westpfahl. 2014. STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In Lori Levin and Manfred Stede, editors, *Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop*, pages 1–10, Dublin, Ireland. Association for Computational Linguistics and Dublin City University. http://www.aclweb.org/anthology/W14-4901.

Heike Zinsmeister, Ulrich Heid, and Kathrin Beck, editors. 2013. *Das STTS-Tagset f/"ur Wortartentagging – Stand und Perspektiven*. Themenheft, Journal for Language Technology and Computational Linguistics 28(1). http://www.jlcl.org.