Adding Value to CMC Corpora: CLARINification and Part-of-Speech Annotation of the Dortmund Chat Corpus

Michael Beißwenger[1], Eric Ehrhardt[2], Andrea Horbach[3], Harald Lüngen[4], Diana Steffen[3], Angelika Storrer[2]

[1] TU Dortmund University, Department of German Language and Literature, D–44221 Dortmund
[2] Mannheim University, Department of German Philology, D–68131 Mannheim
[3] Saarland University, Department of Computational Linguistics and Phonetics, D–66041 Saarbrücken
[4] Institute for the German Language, Department of Central Research: Corpus Linguistics, D–68131 Mannheim

michael.beisswenger@tu-dortmund.de, frehrhar@mail.uni-mannheim.de,
andrea@coli.uni-saarland.de, luengen@ids-mannheim.de,
dsteffen@coli.uni-saarland.de, astorrer@mail.uni-mannheim.de

1 Motivation and Project Framework

ChatCorpus2CLARIN is a curation project of the discipline-specific working group "German Philology" (F-AG 1) within the joint infrastructure project CLARIN-D. In this project, an existing corpus of computer-mediated communication (CMC), the Dortmund Chat Corpus (cf. 2.1), and samples of other CMC resources will be restructured to conform to current standards for the representation of corpora in the Digital Humanities context. The main goal of this work is to pave the way for the inclusion of linguistically annotated CMC resources in CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology. To this end, the project will (1) transform the metadata and the annotations of the chat corpus into a TEI-compliant format, (2) enrich the data by further linguistic annotations, and (3) integrate the resulting resource into the CLARIN-D Corpus Infrastructures at the Institute for the German Language (IDS) and the Berlin-Brandenburg Academy of Sciences (BBAW):

(1) **TEI representation:** For representing the corpus in TEI, the schema drafts and models developed in the TEI special interest group "Computer-mediated communication" are being used. This group is working on a proposal of a TEI standard for CMC genres (Beißwenger et al. 2012, Chanier et al. 2014, Margaretha & Lüngen 2014). In its previous version, the chat corpus has been annotated using a home-grown XML format that describes the main structural features of chat logfiles and user postings as well as selected linguistic phenomena of language use on the internet (emoticons, action words, addressing terms, nicknames). All of these annotations will be transformed into a TEI representation and enriched by additional structural annotations and metadata.

(2) **Additional linguistic annotations:** Except the annotation of selected CMC phenomena, the corpus in its current version does not contain any linguistic annotations. In order to enhance the possibilities for linguistic querying, a layer of part of speech (PoS) annotations will be added to the data. PoS tags using an extended version of the Stuttgart-Tübingen Tagset (STTS, Schiller et al. 1999) have already been added to the corpus using the tools of the project "Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen" (henceforth "Schreibgebrauch", also see http://www.schreibgebrauch.de) developed at Saarland University.

(3) **Integration into CLARIN-D:** The integration of the resource in the CLARIN-D infrastructures comprises its hosting at the CLARIN-D centres BBAW and IDS and its ingestion in the centres' respective repositories for long-term data archiving. It also comprises developing a CMDI representation of metadata for the resource which will be harvestable via OAI-PMH and accessible from the CLARIN VLO (Virtual Language Observatory). The resource will be addressable via PIDs, it will be searchable in the CLARIN-D Federated Content Search and will also be accessible via web services. The conditions of licensing the corpus resource for scientific use will be defined on the basis of a legal expert opinion that is currently being sought. Depending on the outcome of this expert opinion, the Chat Corpus might be licensed with the CLARIN-D end-user license type PUB ("publicly available", cf. Oksanen et al. 2010), ACA-NC (academic, non-commercial use, ibid.), or under an alternative license type like the proposed QAO-NC (use via a query engine that retrieves text passages or KWIC lines the size of citations for users registered in CLARIN-D, cf. Kupietz & Lüngen, 2014).

Our contribution to the NLP4CMC workshop focuses on the subtask of PoS tagging. It describes the goals and work packages of the curation project, the resources, the tagging workflow, and first experiences from the post-processing phase.

2 Resources

2.1 The *Dortmund Chat Corpus*

The Dortmund Chat Corpus (Beißwenger 2013) has been collected at TU Dortmund University. The goal

Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media
pages 12–16, University of Duisburg-Essen, September 28, 2015.

of the corpus project was to create a useful resource for researching the peculiarities and linguistic variation in written computer-mediated communication. The corpus comprises 478 logfile documents with 140 240 user postings or 1M words of German chat discourse representing the use of chat software in different 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 an 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 users. Since 2005, the corpus has been available at http://www.chatkorpus.tu-dortmund.de as an XML version for download and offline querying and as an HTML version for online browsing. It has been widely used as a resource for studying and teaching the peculiarities of German CMC discourse.

2.2 A Tagset for German CMC: 'STTS 2.0'

STTS 2.0 has been created in the context of the DFG scientific network *Empirikom* (http://www.empirikom.net) and of a CLARIN-D initiative and series of workshops (Stuttgart 2012, Tübingen 2013, Hildesheim 2013) for extending the canonical version of STTS (Schiller et al. 1999) for genres which have not been in the scope of the creators of STTS so far (cf. the volume by Zinsmeister et al. 2014). While STTS (1999) focuses mainly on parts of speech in genres of edited text (e.g. newspaper articles, novels), STTS 2.0 builds on the categories of STTS (1999) and extends it with categories and tags for two types of items which have to be taken into consideration when tagging CMC and social media discourse: (1) tags for phenomena which are specific to CMC / social media discourse (emoticons, action words, addressing terms, hash tags, URLs, email addresses), and (2) tags for phenomena which are typical of spontaneous spoken language in colloquial registers (e.g., modal particles, discourse markers, colloquial contractions). These extensions are useful for corpus-based research of both CMC and spoken conversation. A common tag set for phenomena of type (2) will also facilitate the comparison of written CMC with transcripts of spoken conversation.

STTS 2.0 exists in two versions:

- a version described in Bartz et al. (2014) as an intermediate result from and contribution to the discussions in the context of the CLARIN-D STTS initiative 2012/2013. This version has been adopted and slightly modified for adapting a PoS tagger within the project "Schreibgebrauch" at Saarland University in Saarbrücken in 2014/15 (Horbach et al. 2014, henceforth *STTS 2.0-BETA*, cf. 2.2.1);
- a version that builds on Bartz et al. (2014) and includes the results from further discussions in the CLARIN-D STTS initiative and in the Empirikom network and which has been made compatible with the modified STTS defined by Westpfahl & Schmidt (2013) and Westpfahl (2014) for tagging the "Research and Teaching Corpus of Spoken German" (FOLK, http://agd.ids-mannheim.de/folk.shtml) at the IDS Mannheim (Beißwenger et al. 2015, henceforth *STTS 2.0-ALPHA*, cf. 2.2.2).

2.2.1 Tagset Used in the Automatic Annotation Pipeline ('STTS 2.0-BETA')

In order to facilitate and speed up human corpus annotation, we use an automatic tool chain from the "Schreibgebrauch" project to pre-annotate the Dortmund Chat Corpus. The tagging component uses a slightly modified version of the tagsets described in Bartz et al. (2014) and Beißwenger et al. (2015) (cf. Horbach et al. 2014). In particular, the tagset differs in the following points:

- The tagset does not differentiate between ASCII and graphic emoticons.
- The tag for interaction words is split into action word indicators (i.e. the * surrounding the actual interaction word), and the interaction word itself, leading e.g. to tagging results like */AWIND Kaffee/NN trink/AW */AWIND.
- There are no particular tags for various kinds of particles or discourse markers, but they are annotated following the original STTS as adverb or conjunction.
- Extra tags are used to mark words that have been erroneously separated or merged, such as "anzu melden" instead of "anzumelden".

2.2.2 Tagset Used as the Target Tagset in the *ChatCorpus2CLARIN* Project ('STTS 2.0-ALPHA')

STTS 2.0-ALPHA is a slightly revised version of the tagset described in Bartz et al. (2014). It has been described in the guideline document Beißwenger et al. (2015) and will be used as the reference tagset in the Empirikom Shared Task for Automatic Linguistic Annotation of German CMC (https://sites.google.com/site/empirist2015/). It is compatible with the modified STTS that will be used for tagging the FOLK corpus at the IDS (Westpfahl & Schmidt 2013, Westpfahl 2014).

Tab. 1 (see appendix) provides an overview of the tags and categories defined in STTS 2.0-ALPHA. The categories defined for CMC-specific items as well as the extensions for frequent types of colloquial contractions are true extensions to STTS (1999). The categories defined for phenomena which are typical of spontaneously spoken language restructure parts of the categories of STTS (1999). Nevertheless, all modifications and extensions defined in STTS 2.0-

ALPHA result in a category set which is still downwardly compatible with STTS (1999) and therefore allows for interoperability with corpora that have been tagged with STTS (1999) (e.g. DWDS, the "Digital Dictionary of the German Language", http://www.dwds.de).

3 Tagging

The pipeline for pre-annotating the Dortmund Chat Corpus uses tools for sentence segmentation and tokenisation, PoS tagging and lemmatisation. For sentence segmentation and tokenisation we used the open source tokeniser jTok (https://github.com/DFKI-MLT/JTok). It can be adapted to different text types since it uses editable regular expressions to define tokens.

For both PoS tagging and lemmatisation we use the TreeTagger. We employ tagging models from Horbach et al 2014, which have been adapted towards CMC data. In this work, the standard TIGER training data set (Brants et. al. 2004) of about 50 000 newspaper sentences has been extended with relatively small amounts of manually annotated CMC data. they annotated about 12 000 tokens for each of the three CMC genres of forum posts, chat and twitter data with STTS 2.0-BETA tags. The chat subcorpus is taken from the Dortmund Chat Corpus. One third of each dataset has been added to TIGER (boosted 5 times in order to give additional weight to the new material) as training data, while the other two thirds have been held out for testing. These gold annotations can be obtained for research purposes directly from the "Schreibgebrauch" project.

Using a tagger model trained with this enriched training set, performance on the chat part of the test portion of the above mentioned gold-standard annotations could be increased from 71.4% (using an out-of-the-box model trained on TIGER only) to 83.5%. As no lemmatisers adapted towards CMC are available (and our annotations did not comprise lemma information), we used the standard TreeTagger lemmatiser trained on TIGER.

4 Outlook: Post-processing

Parts of the automatically PoS-tagged chat corpus will be manually post-processed, i.e. adapted and amended on the basis of the STTS 2.0-ALPHA tagset as described in section 2.2.2. Post-processing will also concern the levels tokenisation, (orthographic) normalisation, and lemmatisation. The goal of this effort is to create a resource of correct reference annotations for chat data which may be used (a) to demonstrate how a precise tokenisation, PoS annotation, lemmatisation and normalisation of (parts of) a chat corpus will support linguistic users in defining sophisticated corpus queries for their linguistic research questions, (b) as a data set for (re-)training and evaluating NLP tools for the various above-mentioned linguistic processing steps for CMC-specific linguistic items and "non-standard" phenomena in written CMC and social media discourse. Furthermore, the results of the post-processing shall serve as a basis for developing better tokenisation and lemmatisation guidelines for CMC.

Manual post-processing will be carried out by a team of students, using the normalisation editor OrthoNormal in FOLKER ("FOLK-Tools", Schmidt 2012), which has originally been developed and applied for the manual normalisation and correction of POS-tagged spoken language transcripts in the FOLK corpus at the IDS (Westpfahl & Schmidt 2013). A more recent version of FOLKER (preview version 1.2) provided by Thomas Schmidt (IDS) offers a new import and export interface for PoS-tagged ChatXML. Fig. 1 (see appendix) shows a screenshot of editing these data in OrthoNormal. At the NLP4CMC workshop we will present first results of comparing a sample of the automatic PoS annotation using STTS2.0-BETA with an "expert" annotation using STTS2.0-ALPHA and discuss the results by the hand of examples.

References

Books/Papers

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

Bartz, Thomas; Beißwenger, Michael; Storrer, Angelika (2014): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: Journal for Language Technology and Computational Linguistics 28 (1), 157-198. http://www.jlcl.org/2013_Heft1/7Bartz.pdf

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

Beißwenger, Michael; Bartz, Thomas; Storrer, Angelika; Westpfahl, Swantje (2015): Tagset und Richtlinie für das PoS-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline Document, Dortmund 2015. https://sites.google.com/site/empirist2015/home/annotation-guidelines

Brants, Sabine; Dipper, Stefanie; Eisenberg, Peter; Hansen, Silvia; Knig, Esther; Lezius, Wolfgang; Rohrer, Christian; Smith, George; Uszkoreit, Hans (2004): TIGER: Linguistic interpretation of a german corpus. Journal of Language and Computation, Special Issue, 2(4), 597-620.

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

Horbach, Andrea; Steffen, Diana; Thater, Stefan; Pinkal, Manfred (2014): Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. Proceedings of KONVENS 2014, 171-177.

Horbach, Andrea; Thater, Stefan; Steffen, Diana; Fischer, Peter M.; Witt, Andreas; Pinkal, Manfred (2015): Internet Corpora: A Challenge for Linguistic Processing. In: Datenbank-Spektrum 15 (1), 41-47.

Kübler, Sandra; Baucom, Eric (2011): Fast domain adaptation for part of speech tagging for dialogues. In: Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, RANLP, 41-48.

Kupietz, Marc; Lüngen, Harald (2014): Recent developments in DeReKo. In: Nicoletta Calzolari et al. (eds): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland.

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_Heft2/3Margaretha Luengen.pdf

Oksanen, Ville; Lindén, Krister; Westerlund, Hanna (2010): Laundry Symbols and License Management: Practical Considerations for the Distribution of LRs based on experiences from CLARIN. In: Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Malta.

TEI Consortium (2015): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Available online at: http://www.tei-c.org/Guidelines/P5/

Schiller, Anne; Teufel, Simone; Stöckert, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). University of Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmidt, Thomas (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), Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/529_Paper.pdf.

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In Proceedings of the ACL SIGDAT-Workshop, 47-50.

Westpfahl, Swantje; Schmidt, Thomas (2013): POS für(s) FOLK – Part of Speech-Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch. In: Journal for Language Technology and Computational Linguistics 28 (1), 139-156. http://www.jlcl.org/2013_Heft1/6Westpfahl.pdf

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

Zinsmeister, Heike; Heid, Ulrich; Beck, Kathrin Beck (Eds., 2014): Das STTS-Tagset für Wortartentagging - Stand und Perspektiven. Special issue of the Journal for Language Technology and Computational Linguistics. http://www.jlcl.org

Internet Sources

„Analyse und Instrumentarien zur Beobachtung des Schreibgebrauchs im Deutschen“: http://www.schreibgebrauch.de/

CLARIN-D (“Common Language Resources and Technology Infrastructure”) – German section: http://www.clarin-d.de/en/

Dortmund Chat Corpus (“Dortmunder Chat-Korpus"): http://www.chatkorpus.tu-dortmund.de/

DWDS („Digitales Wörterbuch der deutschen Sprache“): http://www.dwds.de

Empirikom (DFG scientific network “Empirische Erforschung internetbasierter Kommunikation”): http://www.empirikom.net

EmpiriST2015 Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication: https://sites.google.com/site/empirist2015/

FOLK (“Forschungs- und Lehrkorpus Gesprochenes Deutsch”): http://agd.ids-mannheim.de/folk.shtml

FOLKER (Transcription editor for FOLK), preview version 1.2 with functionalities for editing data from the Dortmund chat corpus with OrthoNormal: http://agd.ids-mannheim.de/folker.shtml

JTok (rule-based tokeniser): http://heartofgold.opendfki.de/browser/trunk/jtok

TEI special interest group „Computer-mediated communication“: http://www.tei-c.org/Activities/SIG/CMC/

Virtual Language Observatory (VLO): https://vlo.clarin.eu/

Appendix

PoS tag	Category	Examples
I. Tags for phenomena which are specific for CMC / social media discourse:		
EMO ASC	ASCII emoticon	:-) :-(^^ O.O
EMO IMG	Graphic emoticon	
AKW	Interaction word	*lach*, freu, grübel, *lol*
HST	Hash tag	Kreta war super! #urlaub
ADR	Addressing term	@lothar: Wie isset so?
URL	Uniform resource locator	http://www.tu-dortmund.de
EML	E-mail address	peterklein@web.de
II. Tags for phenomena which are typical for spontaneous spoken language in colloquial registers:		
VV PPER	Tags for types of colloquial contractions which are frequent in CMC (APPRART is already existing in STTS 1999)	schreibste, machste
APPR ART		vorm, überm, fürn
VM PPER		willste, darfste, musste
VA PPER		haste, biste, isses
KOUS PPER		wenns, weils, obse
PPER PPER		ichs, dus, ers
ADV ART		son, sone
PTK IFG	'Intensitätspartikeln', 'Fokuspartikeln', 'Gradpartikeln'	sehr schön, höchst eigenartig, nur sie, voll geil
PTK MA	Modal particles	Das ist ja / vielleicht doof. Ist das denn richtig so? Das war halt echt nicht einfach.
PTK MWL	Particle as part of a multi-word lexeme	keine mehr, noch mal, schon wieder
DM	Discourse markers	weil, obwohl, nur, also, ... *with V2 clauses*
ONO	Onomatopoeia	boing, miau, zisch

Tab. 1: Overview of extensions and modifications to STTS (1999) in STTS 2.0-ALPHA (Beißwenger et al. 2015).

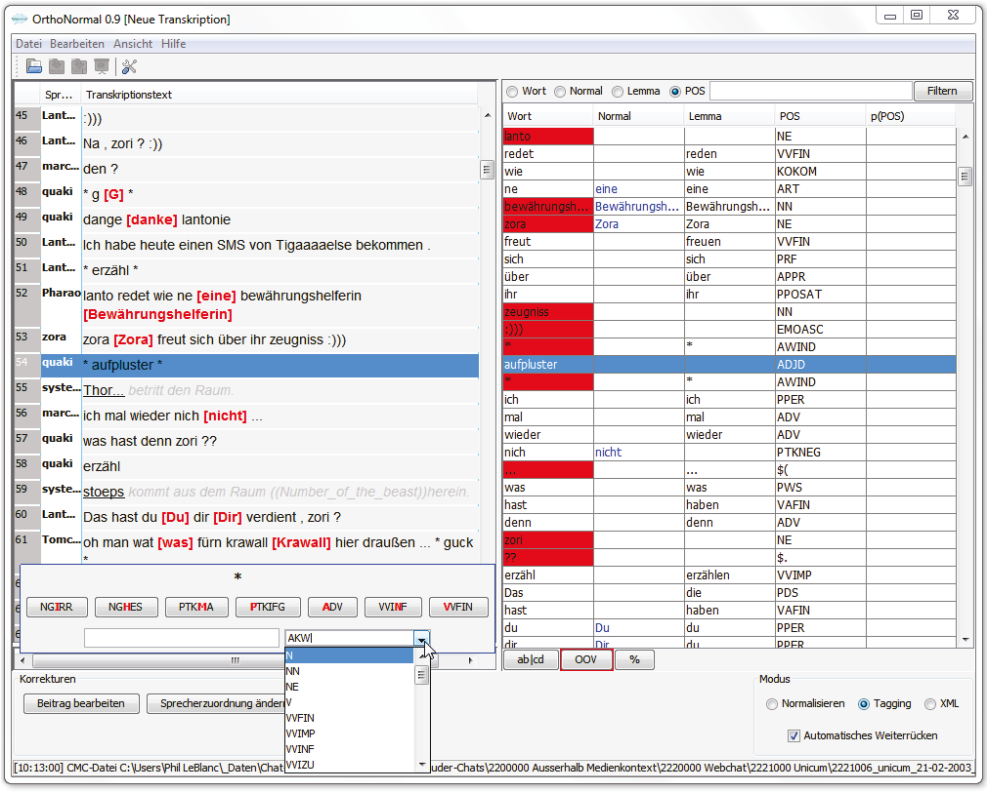

Fig. 1: Editing PoS-tagged ChatXML with OrthoNormal.