Linguistic annotation of social media corpora: To what extent do we have to adapt existing encoding standards and tag sets?

Michael Beißwenger*

* Technical University Dortmund, Germany

Abstract

The talk gives an overview of challenges and open issues in annotating linguistic corpora of social media and computer-mediated communiation (CMC). On the example of an ongoing corpus project in the context of the German CLARIN-D initiative ('ChatCorpus2CLARIN', http://de.clarin.eu/en/curation-project-1-3-german-philology) it presents intermediate results from work dedicated to the modeling and linguistic annotation of CMC. It discusses the question to what extent a modification of existing encoding standards and NLP resources is needed and practical in order to meet two requirements: (1) The resulting schemas and tag sets should allow for an adequate representation of the structural and lingusitic peculiarities of social meda and CMC genres, while at the same time (2) they should not complicate comparative analyses of the language of social media/CMC with the language given in corpora of genres of edited text and of spoken interaction.

In the project ChatCorpus2CLARIN, an existing corpus of German chat communication, the 'Dortmund Chat Corpus' (Beißwenger 2013), and samples of other social media/CMC re-sources 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 inclu-sion of linguistically annotated CMC resources into CLARIN-D corpus infrastructures and to create the prerequisites for investigating linguistic peculiarities of CMC with state-of-the art corpus technology.

The focus of the talk is on the following aspects:
(a) on adapting the encoding guidelines of the Text Encoding Initiative (TEI, http://tei-c.org) for the modeling of structural and linguistic peculiarities of CMC and social media genres,
(b) on adapting a part-of-speech (PoS) tag set for written German (the 'Stuttgat-Tübingen Tagset', Schiller et al. 1999) and using it for adding a layer with part-of-speech annota-tions to the corpus.

The talk will present and discuss a customized TEI schema that has been developed for representing the corpus data and highlight the ideas behind the main modeling decisions, especially with respect to models which are different from TEI-P5 and which have been added or modified in order to capture the peculiarities of CMC.[1].

As a second step, the talk will present and discuss a PoS tag set (Beißwenger et al. 2015) which has been extended with tags for CMC-specific phenomena as well as for phenomena which are particular of spontaneous interactional language (and, thus, for different types of "non-standardness" on the token level, cf. Ljubešić et al. 2015). A basic PoS annotation of the corpus could be achieved by using tagging models developed in the BMBF project "Schreibgebrauch" (http://www.schreibgebrauch.de/) at the University of Saarbrücken (Horbach et al. 2014). For manual post-processing the project uses the editor 'OrthoNormal' in FOLKER (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 Institute for the German Language (IDS) Mannheim (http://agd.ids-mannheim.de/folk.shtml) and which, for use in the ChatCorpus2CLARIN, has been adapted for editing PoS-tagged chat data.

In an outlook, the talk will give an overview of current initiatives and activities in Germany and in the TEI for creating annotation standards for genres of social media / CMC.

[1] Besides the annotation of the corpus resources in the CLARIN-D project, the TEI schema serves as a contribution to the work of the TEI special interest group (SIG) „Computer-mediated communication" which is preparing a proposal for a TEI standard for the representation of CMC. It is based on previous schema versions created and discussed in Beißwenger et al. (2012), Chanier et al. (2014) and Margaretha/Lüngen (2014). The schema and its documentation will be made available in the form of an ODD document on the SIG pages in the TEI wiki as of October, 23): http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication

References

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: Zeitschrift für germanistische Linguistik 41 (1), 161-164. Extended version: http://tinyurl.com/chatkorpus

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

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.

Ljubešić, Nikola; Fišer, Darja; Erjavec, Tomaž; Čibej, Jaka; Marko, Dafne; Pollak, Senja; Škrjanec, Iza (2015): Predicting the Level of Text Standardness in User-generated Content. In: Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria, Sep 7–9 2015, 371–378, http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf

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

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.