Integrating corpora of computer-mediated communication into the language resources landscape: Initiatives and best practices from French, German, Italian and Slovenian projects

Michael Beißwenger
University of Duisburg-Essen, Germany
michael.beisswenger@uni-due.de

Thierry Chanier
Université Blaise Pascal, France
thierry.chanier@univ-bpclermont.fr

Isabella Chiari
Sapienza Università di Roma, Italy
isabella.chiari@uniroma1.it

Tomaž Erjavec
Jožef Stefan Institute
Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Darja Fišer
University of Ljubljana
Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Axel Herold
Berlin-Brandenburg Academy of Sciences, Berlin, Germany
herold@bbaw.de

Nikola Ljubešić
Jožef Stefan Institute
Ljubljana, Slovenia
nikola.ljubesic@ffzg.hr

Harald Lüngen
Institute for the German Language, Mannheim, Germany
luengen@ids-mannheim.de

Céline Poudat
Université de Nice Sophia Antipolis, France
poudat@unice.fr

Egon Stemle
EURAC, Bolzano
Italy
egon.stemle@eurac.edu

Angelika Storrer
University of Mannheim, Mannheim, Germany
astorrer@mail.uni-mannheim.de

Ciara Wigham
Université Blaise Pascal, France
ciara.wigham@univ-bpclermont.fr

Abstract

The paper presents best practices and results from projects in four countries dedicated to the creation of corpora of computer-mediated communication and social media interactions (CMC). Even though there are still many open issues related to building and annotating corpora of that type, there already exists a range of accessible solutions which have been tested in projects and which may serve as a starting point for a more precise discussion of how future standards for CMC corpora may (and should) be shaped like.

Introduction

The paper presents best practices and results from projects in four countries dedicated to the creation of corpora of computer-mediated communication and social media interactions (henceforth referred to as *CMC*). The projects relate to each other via a bottom-up network of researchers interested in fostering the transfer of expertise and solutions for handling this relatively new type of language resources and for modeling the structural and linguistic peculiarities of (written and multimodal) discourse that is found in chat, forum, sms and whatsapp interactions, in weblogs and wikis, on social network sites or in multimodal CMC environments. This new type of discourse exhibits features that cannot be adequately handled through the application of schemas, categories and tools which have been developed for the representation, annotation and processing of discourse which conforms to the

written standard and the structural conventions of established text types (e.g., newspaper articles, prose, scientific articles). In addition, the collection and redistribution of CMC data in linguistic corpora faces corpus creators with legal and ethical issues which are not yet completely covered by existing laws and ethical standards. Last but not least there are no established standards for metadata and for the documentation of the (technological, hypermedial and social) context in which CMC data are typically embedded, produced and used.

Corpus-linguistic approaches to CMC have so far not found answers to all of these challenges. Nevertheless, existing projects in the field have created an encouraging range of results and best practices which have been tested with existing corpora and which are worth further discussions. The joint goal of the projects described in this paper is to pave the ground for future standards which will allow CMC corpora to be interoperable (a) with each other and (b) with language resources for other types of discourse (text and speech corpora).

In the following sections we describe existing initiatives for the development of standards and for the exchange of knowledge related to the collection and representation of CMC corpora (Sect. 2) and give an overview of results and best practices from CMC corpus projects in four countries which may be useful for other projects in the field and which may serve as a starting point for a more precise discussion of how future standards for CMC corpora may (and should) be shaped like (Sect. 3).

Initiatives and previous cooperation in the field

Since 2013 a loose network of projects with a joint interest in building and annotating CMC corpora has set up two initiatives in order to (1) strengthen the exchange of expertise and best practices between projects and (2) lead the discussion of a representation standard for CMC genres in the context of a well-acknowledged standardization initiative in the Digital Humanities:

- The network has established a series of international workshops and conferences dedicated to the creation of CMC corpora with previous events held in Dortmund/DE (2013, 2014), Rennes/F (2015) and Ljubljana/SI (2016). Since 2015 these conferences are defined as peer-reviewed international events with a coordinating and a scientific committee (*Conference on CMC and Social Media Corpora for the Humanities*, http://www.cmc-corpora.org).
- The network succeeded with a proposal for the creation of a new special interest group (SIG) on Computer-Mediated Communication in the *Text Encoding Initiative* (*TEI*, http:///tei-c.org) in 2013. The goal of this SIG is to extend the TEI framework with additions dedicated to the representation of the structural and linguistic peculiarities of CMC genres. Starting from a discussion of a first schema draft defined by Beißwenger et al. (2012) the SIG created two advanced schema drafts ('CoMeRe schema', 2014, 'CLARIN-D schema', 2015) which have been tested with French and German corpora and which are currently being adopted by further projects.

Overview of projects and best practices from the *cmc-corpora* network

3.1 France

The *CoMeRe* project ('Communication Médiée par les Réseaux', 2013-2015, National Consortium on Written Corpora, *TGIR Huma-Num*) brought together researchers who had previously collected different types of CMC corpora in their local research teams or in previous research projects, and had structured these in a variety of formats (different XML schemas for text chat corpora, SMS corpora, for LEarning and TEaching Corpora). The primary aim of the project was to design a common model for CMC discourse that would fit the pre-existing CMC corpora, as well as new corpora collected both during the project or post-project. The secondary aim was to release these corpora in a common repository as open data, in order to provide access to a dataset with significant coverage to colleagues working within the field of discourse analysis and who are interested in the linguistic study of CMC genres. To address the project's primary goal, it was first necessary to settle on a common document model that would fit different types of multimodal CMC data. Rather than working genre by genre, the project members initially worked within the abstract concept of the 'Interaction Space'. This concept is detailed in Chanier et al. (2014). Briefly, an Interaction Space is located within a timeframe, during

which interactions occur between a set of participants within an online location. This location is defined by the properties of the set of environments used by the participants who may be either individual members or groups. The environments may be synchronous or asynchronous, mono- or multimodal, simple or complex. The traces of actions within an environment and one particular modality are termed an 'act'. Working from this concept, a TEI schema was proposed ('CoMeRe schema', 2014). Relationships between the Interaction Space definitions of the environment, found in the <teiHeader>, and its actual use by participants in interactions, described in the <body> part of the TEI file, appear through the attribute @type of the <post> element. All 14 corpora in the CoMeRe repository (2016) have been structured in this manner.

3.2 Germany

First ideas towards an annotation schema for CMC in Germany have been discussed in the context of the DFG scientific network *Empirikom* (http://www.empirikom.net) and together with the concept for a reference corpus of German CMC (*DeRiK*, Beißwenger et al., 2013). More recently, the idea to create one large reference corpus comprising several CMC genres developed into the concept of a collection of corpora which are annotated on the basis of the same schema. First corpora of that type are the "CLARINified" version of the Dortmund Chat Corpus (Lüngen et al., 2016), the Wikipedia Corpus (Margaretha and Lüngen, 2014) and the Usenet Corpus in DEREKO (Schröck and Lüngen, 2015). Further resources are in preparation. In the context of the CLARIN-D curation project *ChatCorpus2CLARIN* schema drafts from the TEI CMC-SIG (cf. Sect. 2) have been further developed ('CLARIN-D schema', 2015) and used for the representation of a 1 million token chat corpus with part-of-speech annotations. In addition, the project initiated writing of a legal opinion on republishing CMC data as part of the CLARIN corpus infrastructure (iRights, 2016). Linguistic annotation was done partially automatically on the basis of an extended version of the *STTS* tag set which is a de-facto standard for German text corpora. The extensions to this tag set include tags for typical phenomena of CMC discourse; the tag set as a whole ('STTS 2.0', Beißwenger et al., 2015) is downwardly compatible with the "standard" STTS and with the STTS version used for PoS tagging the FOLK corpus of spoken language at IDS Mannheim (Westpfahl and Schmidt, 2016). STTS 2.0 was originally developed as a resource for the community shared task *EmpiriST2015* which was designed to foster the adaptation of NLP tools to the linguistic peculiarities of German CMC (Beißwenger et al., 2016). All resources from EmpiriST2015 are available on the web (http://sites.google.com/site/empirist2015/). The TEI version of the PoS-annotated chat corpus will be integrated into the CLARIN-D corpus infrastructures.

3.3 Italy

In the *DiDi* project at the EURAC, Bozen, Frey et al. (2015) created a CMC corpus of South Tyrolean German from Facebook (FB) users' wall posts, comments on wall posts and private messages, supplemented with user-provided socio-demographic data. All FB data was automatically annotated with language codes, and manually normalized and anonymized. Semi-automatic token level annotations include part-of-speech tags and CMC phenomena (e.g. emoticons, emojis, and iteration of graphemes and punctuation). The corpus was collected with an integrated web strategy for mixed sociolinguistic research methodologies in the context of social media corpora, and also resulted in recommendations for collecting private, non-public CMC data. With a meticulously crafted procedure where users had to explicitly agree with the distribution of their data and were also able to restrict the collection of data to the publicly available part, the anonymized corpus without the private messages is now freely available for researchers, and the complete anonymized corpus is available after signing an agreement.

Web2Corpus ("Corpus italiano di comunicazione mediata dal computer") is a project funded by Sapienza University of Rome in 2011 aimed at investigating meaning negotiation strategies in CMC (http://www.glottoweb.org/web2corpus/). It focuses on conversational, interactive, public, written communication in order to build a genre-balanced CMC corpus of the Italian language. The genres included are forums, blogs, newsgroups, social networks and chats (Chiari and Canzonetti, 2014). The corpus has been fully anonymized (by masking) and XML-annotated both for macro-structural properties (thread, post, sender details, subject, date, time, links and embedded media, web action elements and CMC-specific emoticons and tags and addressing terms, etc.). The anonymization

guidelines will soon be released in the form of best practices taking into account current legislation and technical solutions. At present the corpus is being processed linguistically with a statistical PoS tagger and lemmatizer, including a reference machine dictionary (Common Lexicon of Italian) developed in order to include CMC-specific lexical items, and will be subsequently manually checked. A first qualitative and quantitative analysis is presented in Chiari (ed., 2016 in press).

3.4 Slovenia

The *Janes* project (http://nl.ijs.si/janes/) is compiling a corpus of Slovene (Erjavec et al., 2015a) that contains tweets, forum posts, news comments, blogs and blog comments, and user and talk pages from Wikipedia. The current version v0.4 contains around 9 million posts comprising 200 million tokens. In addition to the metadata obtained during crawling or from the text, such as date and time of posting, username, post URL, discussion thread, and the number of likes and retweets, the corpus is also annotated with metadata that were added manually, such as user type and gender, or automatically, such as user's region, text sentiment and standardness level. The standard linguistic annotation workflow has been adapted to better tackle CMC-specific features and comprises tokenization, sentence segmentation, rediacritisation, normalization, morphosyntactic tagging and lemmatization (Ljubešić and Erjavec, 2016). The corpus is encoded in bespoke XML that closely reflects the structure of the corpus and all its metadata. Version 1 will be encoded in a CMC-aware TEI proposed by Beißwenger et al. (2012). Apart from the XML source files, the corpus is also made available to linguists on the local installation of the noSketchEngine and SketchEngine concordancers (Rychlý, 2007), both as the entire Janes v0.4 corpus with the metadata that all the subcorpora have in common and as separate subcorpora with all the metadata available for the given subcorpus. Access to the corpus is currently restricted to project members but steps are being taken to comply with the copyright, terms of use and privacy issues in order to make an annonymised, sampled and shuffled corpus available to other researchers as well by the end of the project (Erjavec et al., 2015b).

Outlook

In this paper we gave an overview of first results and best practices from projects in four countries dedicated to the creation of corpora of computer-mediated communication and social media interactions (CMC). The joint goal of the projects is to establish standards for the collection and representation of CMC corpora and for their integration into common resources infrastructures.

Up to now the network has brought forward two main initiatives: a conference series dedicated to all issues related to building and annotating CMC corpora and a TEI-SIG focused on the integration of standards for CMC resources into the TEI framework. Both initiatives are "bottom up" approaches with the goal to connect researchers all over Europe and to work on solutions driven by practices that have proven useful in ongoing projects. The latest edition of the conference included 22 contributions by 40 authors from 24 research institutions in 11 countries (Fišer and Beißwenger, 2016).

Nevertheless, there's still a lot of open, non-trivial issues in the field. One example is the lack of legal standards for collecting and republishing CMC data as part of language resources. Corpus builders are typically laymen when it comes to legal issues. A general legal opinion on these issues commissioned and disseminated by and via an acknowledged language resources initiative would therefore be an important prerequisite for the further development of the CMC corpora landscape and community.

In view of the importance of CMC in everyday communication, in business, public administration, science and education, efforts in the field of establishing state-of-the-art research resources infrastructures for the analysis of CMC phenomena are an investment in our future knowledge about how the adoption of CMC technologies affects society and how communicative practices reflect the presence of CMC as an innovative means for the organization of social interaction.

References

[Beißwenger et al., 2012] 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*, (3) (doi: 10.4000/jtei.476). http://jtei.revues.org/476

[Beißwenger et al., 2013] Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. 2013. DeRiK: A German Reference Corpus of Computer-Mediated Communication. *Literary and Linguistic Computing*, 28(4):531-537. (doi: 10.1093/llc/fqt038).

[Beißwenger et al., 2015] Michael Beißwenger, Thomas Bartz, Angelika Storrer, and Swantje Westpfahl (2015b): *Tagset and guidelines for the PoS tagging of language data from genres of computer-mediated communication / social media*. http://sites.google.com/site/empirist2015/home/annotation-guidelines.

[Beißwenger et al., 2016] 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. Berlin, Germany, 44–56. http://aclweb.org/anthology/W/W16/W16-2606.pdf

[Chanier et al., 2014] Thierry Chanier, Celine 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

[Chiari, 2016 in press] Isabella Chiari (ed). 2016 in press. *Capirsi e fraintendersi al computer. La negoziazione del senso nella conversazione sui nuovi media*, Carocci, Roma.

[Chiari and Canzonetti, 2014] Isabella Chiari and Alessio Canzonetti. 2014. *Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione*. In: Enrico Garavelli and Elina Suomela-Härmä (eds.). Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua. Atti del XII Convegno della Società Internazionale di Linguistica e Filologia Italiana (SILFI), Helsinki 18-19 June 2012. Franco Cesati Editore, Firenze, 595-606.

[CLARIN-D schema, 2015] CLARIN-D TEI schema for CMC corpora. 2015. http://wiki.tei-c.org/index.php?title=SIG:CMC/clarindschema

[CoMeRe, 2016] CoMeRe repository. 2016. *Corpora of Computer-Mediated Communication in French.* http://hdl.handle.net/11403/comere

[CoMeRe schema, 2014] CoMeRe TEI RelaxNG XML schema for CMC corpora, version 2. 2014. https://repository.ortolang.fr/api/content/comere/v2/tei_cmr.rng

[Erjavec et al., 2015a] Tomaž Erjavec, Darja Fišer, and Nikola Ljubešić. 2015a. *Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes.* Fišer, Darja (ed.). Zbornik konference Slovenščina na spletu in v novih medijih. Ljubljana, Znanstvena založba Filozofske fakultete, 20–26.

[Erjavec et al., 2015b] Tomaž Erjavec, Jaka Čibej, and Darja Fišer. 2015b. *Pravna podlaga za zagotavljanje prostega dostopa korpusov spletnih besedil*. Smolej, Mojca (ed.). OBDOBJA 34: Slovnica in slovar – aktualni jezikovni opis. Ljubljana, Znanstvena založba Filozofske fakultete, 193–199.

[Fišer and Beißwenger, 2016] Darja Fišer and Michael Beißwenger (eds.). 2016. *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora2016)*. University of Ljubljana: Faculty of Arts. http://nl.ijs.si/janes/cmc-corpora2016/proceedings/

[Frey et al., 2015] Jennifer-Carmen Frey, Aivars Glaznieks, and Egon W. Stemle. 2015. *The DiDi Corpus of South Tyrolean CMC Data*. In: Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media at GSCL2015 (NLP4CMC2015), Essen, Germany, 1–6. https://sites.google.com/site/nlp4cmc2015/program

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

[Ljubešić and Erjavec, 2016] Nikola Ljubešić and Tomaž Erjavec. 2016. *Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: The Case of Slovene*. In: Proceedings of Language Resources and Evaluation Conference (LREC) 2016. Portorož, Slovenia, 1527–1531.

[Lüngen et al., 2016] Harald Lüngen, Michael Beißwenger, Eric Ehrhardt, Axel Herold, and Angelika Storrer. 2016. *Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN*. In: Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), 156–164. https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf

[Margaretha and Lüngen, 2014] Eliza Margaretha and Harald Lüngen. 2014. Building Linguistic Corpora from Wikipedia Articles and Discussions. *Journal of language Technology and Computational Linguistics*, 29 (2):59–82. http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf

[Schröck and Lüngen, 2015] 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). Essen, Germany, 17-22. https://sites.google.com/site/nlp4cmc2015/program

[Westpfahl and Schmidt2016] Swantje Westpfahl and Thomas Schmidt. 2016. *FOLK-Gold – A GOLD standard for Part-of-Speech- Tagging of Spoken German*. In: Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC16), Paris, France.