Connecting Resources: Which Issues Have to be Solved to Integrate CMC Corpora from Heterogeneous Sources and for Different Languages?

Michael Beißwenger[1], Ciara Wigham[2], Carole Etienne[3], Darja Fišer[4], Holger Grumt Suárez[5], Laura Herzberg[6], Erhard Hinrichs[7], Tobias Horsmann[1], Natali Karlova-Bourbonus[5], Lothar Lemnitzer[8], Julien Longhi[9], Harald Lüngen[10], Lydia-Mai Ho-Dac[11], Christophe Parisse[12], Céline Poudat[13], Thomas Schmidt[10], Egon Stemle[14], Angelika Storrer[6], Torsten Zesch[1]

[1] University of Duisburg-Essen, Germany [2] University Clermont-Auvergne, France [3] ICAR Laboratory Lyon, France
[4] University of Ljubljana, Slovenia [5] Justus-Liebig-Universität Gießen, Germany [6] University of Mannheim, Germany
[7] Eberhard-Karls-Universität Tübingen, Germany [8] Berlin-Brandenburg Academy of Sciences, Germany
[9] Université de Cergy-Pontoise, France [10] Institute for the German Language, Mannheim, Germany
[11] Université Toulouse 2, France [12] Université Paris Nanterre, France
[13] Université Nice Côte d'Azur, France [14] Eurac Research, Bolzano, Italy

michael.beisswenger@uni-due.de, ciara.wigham@uca.fr, carole.etienne@ens-lyon.fr, darja.fiser@ff.uni-lj.si,
Holger.H.Grumt-Suarez@germanistik.uni-giessen.de, lherzber@mail.uni-mannheim.de,
erhard.hinrichs@uni-tuebingen.de, tobias.horsmann@uni-due.de, Natali.Karlova-Bourbonus@zmi.uni-giessen.de,
lemnitzer@bbaw.de, julien.longhi@u-cergy.fr, luengen@ids-mannheim.de, hodac@univ-tlse2.fr,
cparisse@u-paris10.fr, celine.poudat@unice.fr, thomas.schmidt@ids-mannheim.de,
egon.stemle@eurac.edu, astorrer@mail.uni-mannheim.de, torsten.zesch@uni-due.de

The paper reports on the results of a scientific colloquium dedicated to the creation of standards and best practices which are needed to facilitate the integration of language resources for CMC stemming from different origins and the linguistic analysis of CMC phenomena in different languages and genres. The key issue to be solved is that of interoperability – with respect to the structural representation of CMC genres, linguistic annotations metadata, and anonymization/pseudonymization schemas. The objective of the paper is to convince more projects to partake in a discussion about standards for CMC corpora and for the creation of a CMC corpus infrastructure across languages and genres. In view of the broad range of corpus projects which are currently underway all over Europe, there is a great window of opportunity for the creation of standards in a bottom-up approach.

Keywords: corpora, research infrastructures, annotation, anonymization

1. Background and Motivation

The paper reports on the results of a scientific colloquium (https://sites.google.com/view/dhcmc2017/) dedicated to the creation of standards and best practices which are needed as a prerequisite for the exchange, interconnection, and combined analysis of CMC corpora of different origins, for different languages and different genres. The goal of the colloquium which was held with funding from the French Embassy in Germany was to determine open issues which have to be solved to represent CMC corpus data (including metadata and annotations) using interoperable formats. From a wider perspective, the colloquium addressed not only issues of interoperability of CMC corpora between one another but also the interoperability of CMC corpora with corpora of other types, namely text corpora and spoken language corpora.

To make the goal of interoperability and of the development of standards more tangible for the community of CMC researchers and corpus creators, the colloquium outlined the scenario of creating a multilingual and genre-heterogeneous demo corpus that would include samples from existing CMC corpora in different languages and on different CMC genres and which would also present CMC in the context of other discourse domains through the inclusion of samples from text and spoken language corpora.

In the following sections, we summarize results and remaining open issues towards the creation of standards and towards an interoperability of corpora as were determined during the colloquium. The overview is based on input from representatives of the following corpus projects and language resource infrastructure projects:

a) *CMC corpora:*

- *CoMeRe:* a collection of 14 French corpora for 9 different CMC genres (including multimodal genres) represented in TEI (Chanier et al., 2014) and available for download (CC BY, OpenData) via ORTOLANG. (Longhi and Wigham, 2015)[1]
- *DEREKO-News:* Corpus of German Newsgroups in DEREKO, since 2013, 98 million tokens (Schröck and Lüngen, 2015).[2]
- *DEREKO-Wikipedia:* Wikipedia corpora in DEREKO: German language article, talk and user talk (Margaretha and Lüngen, 2014), 581 million tokens, available for online querying via COSMAS II.
- *DiDi corpus:* The CMC corpus from the DiDi project with 570.000 tokens of German, Italian and South Tyrolean Facebook posts and interactions, available

[1] http://hdl.handle.net/11403/comere
[2] https://cosmas2.ids-mannheim.de/

for online querying via ANNIS (Frey et al., 2016).[3]

- *Dortmund Chat Corpus 2.0:* corpus of German chat discourse represented in TEI, available as part of the CLARIN-D corpus infrastructure (Lüngen et al., 2016, Beißwenger et al., 2017).
- *DWDS blog corpus:* The blog corpus in the corpus collection of the DWDS project: 103 million tokens from CC-licensed, mainly German blog entries, available for online querying (Barbaresi, 2016).[4]
- *Gießen scienceblog corpus:* ongoing project at the University of Gießen; goal: creation and annotation of a corpus of German science blogs. (Grumt Suárez et al., 2016)
- *Janes corpus:* The Corpus of Nonstandard Slovene comprising >200 million tokens from tweets, forum posts, blogs, comments on news articles and Wikipedia discussions (Fišer et al., 2016, 2017).[5]
- *MoCoDa:* ongoing project at University of Duisburg-Essen (2017–), goal: creation of a database with a web frontend for repeated, donation-based collection of mobile CMC (whatsapp, sms & co.).
- *Wikiconflits:* TEI-CMC-encoded corpus of French Wikipedia talk pages associated to conflicts capturing 7 topics related to (pseudo-)science with 4456 posts (Poudat et al., 2017).
- *WikiTalk:* TEI-P5-encoded corpus of French Wikipedia talk pages (365,612 pages, >1M threads).
- *DiscoWiki:* corpus of an ongoing project with the goal of annotating relevant characteristics for conflict detection and description at the thread level; corpus based on a selection of talk pages extracted from *Wikiconflits* and *WikiTalk* (Ho-Dac and Laippala, 2017).

b) *Corpora of other types:*

- Text corpus collection at the BBAW, Berlin (DWDS corpora) (Geyken et al., 2017).
- German reference corpus at the IDS Mannheim (DEREKO) (Lüngen, 2017).
- The French spoken language corpora *Colaje* (Morgenstern and Parisse, 2012) and *Orfeo*[6].
- Research and Teaching Corpus of Spoken German (FOLK) at the IDS Mannheim (Schmidt, 2016).[7]

c) *Corpus infrastructure projects:*

- *CLARIN-D*, the German national branch of the Eurpoean CLARIN initiative.[8]
- *ORTOLANG*, the French infrastructure for Open Resources and TOols for LANGuage.[9]

The objective of presenting this summary at CMC-Corpora17 is to convince more people with corpus projects in the field to join the discussion about standards for CMC corpora and – probably – for transforming the idea of a CMC demo corpus into a cooperative project with the participation of a broad range of projects and researchers.

2. State of the art and open issues

2.1 Basic representation format

Since 2012, the special interest group (SIG) "computer-mediated communication" in the Text Encoding Initiative (TEI) has created three TEI extensions for the representation of CMC data and tested these extensions with different CMC genres and corpora for French and German (Beißwenger et al., 2012, Chanier et al., 2014, Lüngen et al., 2016). All three extensions are available in the form of RNG schemas and ODD documents and ready to be used for annotation in other projects.[10] Current work of the SIG is focused on the transformation of the available extensions into a "feature request" which is necessary to make an official suggestion for extending the TEI guidelines with models for CMC. Discussions at the colloquium showed that for further dissemination of the TEI extensions for CMC is desirable

- to document practices, tools, and guideline documents from projects that have already converted raw data into TEI and make them available as Open Access resources to facilitate the conversion of corpus data into TEI for colleagues who have not worked with TEI before;
- to diffuse information/documentation concerning toolchains that can be used on CMC data as currently there is little support for users to help them process resources in one way or another.

One possible outlet for the diffusion of resources of this type would be the CLARIN Language resource switchboard (Zinn, 2016) that is currently being developed within CLARIN-PLUS as a means to link linguistic resources with the tools that can process them. It aims to create a single point of access where users can find the tools that fit their needs and their language resource.

2.2 Natural language processing

Different creators of CMC corpora are using a different types of linguistic annotations and different (e.g., language-specific) tagsets. For mapping annotations in existing corpus resources without the need to perform a complete re-annotation, the resources of the Universal Dependencies Initiative (UD) [11] may provide CMC researchers with "a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary". Nevertheless, UD does not provide any tags for the

[3] http://www.eurac.edu/didi
[4] https://www.dwds.de
[5] http://nl.ijs.si/janes/
[6] http://www.projet-orfeo.fr/
[7] http://agd.ids-mannheim.de/folk.shtml
[8] https://www.clarin-d.de/de/
[9] http://ortolang.fr
[10] These resources are available via the SIG space in the TEI wiki: https://wiki.tei-c.org/index.php?title=SIG:CMC
[11] http://universaldependencies.org/introduction.html

description of CMC-specific phenomena. In order to determine the feasibility of mapping tagset extensions used in CMC corpora in different languages onto each other, the group agreed that it will be helpful to compare how CMC-specific phenomena are treated in the DiDi, CoMeRe, CLARIN-D and Janes (and other CMC) corpora and check the extent to which they are already compatible with each other / analyse the effort needed to transform them into a compatible structure. This could be the subject of a workshop or short-term project in the near future. Further work on this topic should also take into account the tagsets and resources both from the PoSTWITA shared task on PoS tagging Italian social media data[12] and from the EmpiriST shared task on PoS tagging of German CMC and web corpora data[13].

2.3 Anonymization

In order for CMC corpora to conform to the restrictions of national data protection rights (DPR), existing corpora projects have developed different strategies and practices for removing or masking personal data. From a DPR perspective, it is required that the data included in corpora which are made available to the scientific community should be represented in a way that the expenses that one would have to invest to identify a certain person are so high that it seems unrealistic that anybody would invest them. Different annymization approaches have been adopted by the different projects. It may be a fruitful topic for further investigation (i.e. in form of a workshop or short-term project) to compare, in detail, the results of anonymization vs. pseudonymization approaches adopted in different projects to determine the best balance between preserving as much of the semantics of the original data as possible (which is an important resource especially for qualitative analyses) while on the other hand staying as feasible as possible in terms of time-cost factors. One action point could be to make anonymization guidelines from different projects available to other colleagues so that procedures employed can be re-used. Examples for best practices for anonymizing CMC corpora have been developed and employed in the DiDi project (Frey et al., 2015), in the CLARIN-D curation project ChatCorpus2 CLARIN (Lüngen et al., 2017) and in CoMeRe (Chanier and Jin, 2013).

2.4 Metadata

The creation and representation of metadata for CMC corpora and for single interactions preserved in them is a huge and urgent open issue. One key point is for metadata to include information about the version of the communication platform, given the rate at which communication platforms evolve. In the future, should access to older version no longer be available, at least prose descriptions of the communication platform at the time of data collection would be practical for future corpora end-users and mandatory to guarantee corpus data sustainability. A comparison of the metadata captured in different corpus projects (be it in the form of annotations, be it in the form of prose descriptions) could be a fruitful contribution to a more precise discussion of (i) what types of metadata are needed in and for CMC corpora, (ii) to what extent these metadata are specific to individual CMC genres or for CMC in general, (iii) the preservation of which types of metadata could be standardized, and (iv) how a basic representation schema for CMC metadata (e.g., in the TEI Header) could look like.

3. Outlook

In view of the broad range of corpus projects which are currently underway all over Europe (Beißwenger et al., 2017a), there is a great window of opportunity for the creation of standards for CMC corpora in a bottom-up approach. The discussions obtained on this issue at CMCCorpora17 shall be included in the creation of a white paper giving a more precise outline of future work for the issues addressed in this paper. The creation of a demo corpus including samples from different existing CMC corpora could support the further investigation of open issues and provide valuable feed-back for existing best practices in the field. A prerequisite would be a "critical mass" of resources and researchers who are willing to contribute to the creation of such a corpus. As a first step of preparatory work it is planned to set up a platform for the exchange of tools, tips and case studies between projects in order to facilitate the dissemination of knowledge and best practices.

4. References

Barbaresi, A. (2016). Efficient construction of metadata-enhanced web corpora. In *Proceedings of the 10th Web as Corpus Workshop, Association for Computational Linguistics*, pp. 7-16. https://hal.archives-ouvertes.fr/hal-01371704v2/document.

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* 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Beißwenger, M., Lüngen, H., Schallaböck, J., Weitzmann, J.H., Herold, A., Kamocki, P., Storrer, A., Wildgans, J. (2017, in press). Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens. In: M. Beißwenger (Ed.), *Empirische Erforschung internetbasierter Kommunikation*. Berlin/New York: de Gruyter (Empirische Linguistik / Empirical Linguistics).

Beißwenger, M., Chanier, T., Erjavec, T., Fišer, D., Herold, A., Lubešic, N., Lüngen, H., Poudat, C., Stemle, E., Storrer, A., Wigham, C. (2017a). Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In: L. Borin (Ed.), *Selected papers from the CLARIN Annual*

[12] http://corpora.ficlit.unibo.it/PoSTWITA/index.php?slab=guidelines

[13] Tagset/guidelines: https://sites.google.com/site/empirist2015/; results: WAC-X/EmpiriST (2016).

Conference 2016, Aix-en-Provence, 26–28 October 2016 (Linköping University Electronic Conference Proceedings 136), pp. 1-18. http://www.ep.liu.se/ecp/contents.asp?issue=136

Chanier, T., Jin, K. (2013). *Defining the online interaction space and the TEI structure for CoMeRe corpora. Projet CoMeRe (Communication Médiée par les Réseaux).* https://corpuscomere.files.wordpress.com/2014/01/tei-cmc-comere-interactionspace_131231.pdf

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. *Journal of language Technology and Computational Linguistics*, 29(2), pp. 1–30. http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf.

Fišer, D., Erjavec, T., Ljubešić, N. (2017). The compilation, processing and analysis of the Janes corpus of Slovene user-generated content: In C.R. Wigham, G. Ledegen (Eds.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse*. Paris: L'Harmattan (Humanités numériques), pp. 125–138.

Fišer, D., Erjavec, T., Ljubešić, N. (2016). JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0*, 4(2), pp. 67–99.

Frey, J.C., Glaznieks, A., Stemle, E.W. (2015). The DiDi Cor-pus of South Tyr-olean CMC Data. In *Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media (NLP4CMC2015), Essen, Germany*.

Frey, J.-C., Glaznieks, A., Stemle, E. (2016). The DiDi Corpus of South Tyrolean CMC Data: A Multilingual Corpus of Facebook Texts. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*. ceur-ws.org/Vol-1749/paper27.pdf.

Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., Lemnitzer, L. (2017, in press). Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache“ (DWDS). *Zeitschrift für germanistische Linguistik*, 45 (2).

Grumt Suárez, H., Karlova-Bourbonus, N., 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 (cmc-corpora2016), Ljubljana, Slovenia*. http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Grumt_et_al_Compilation-and-Annotation.pdf.

Ho-Dac, L.-M., Laippala, V. (2017). Le corpus WikiDisc, une ressource pour la caractérisation des discussions en ligne. In C.R. Wigham, G. Ledegen (Eds.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse.* Paris: L'Harmattan (Humanités numériques), pp. 107–124.

Longhi, J., Wigham, C.R. (2015). *Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow.* Poster at Corpus Linguistics 2015, Lancaster, United Kingdom. https://halshs.archives-ouvertes.fr/halshs-01176061.

Lüngen, H. (2017). DeReKo – Das Deutsche Referenzkorpus. Schriftkorpora der deutschen Gegenwartssprache am Institut für Deutsche Sprache in Mannheim. *Zeitschrift für germanistische Linguistik*, 45 (1), pp. 161–170.

Lüngen, H., Beißwenger, M., Ehrhardt, E., Herold, A., Storrer, A. (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), Bochum, Germany*, pp. 156–164. https://www.linguistics.rub.de/konvens16/pub/20_ konvensproc.pdf.

Lüngen, H., Beißwenger, M., Herzberg, L., Pichler, C. (2017). Anonymisation of the Dortmund Chat Corpus 2.1. In *Proceedings of the 5th Conference on CMC and Social Media corpora for the Humanities, Bolzano, Italy.*

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

Morgenstern, A. & Parisse, C. (2012). The Paris corpus. *Journal of French Language Studies*, 22, pp. 7–12.

Poudat, C., Grabar, N., Paloque-Berges, C., Chanier, T., Juin, K. (2017). Wikiconflits: un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In C.R. Wigham, G. Ledegen (Eds.), *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse.* Paris: L'Harmattan (Humanités numériques), pp. 19–36.

Schmidt, T. (2016). Good practices in the compilation of FOLK, the research and teaching corpus of spoken German. *International Journal of Corpus Linguistics*, 21(3), pp. 396–418.

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), Essen, Germany*, pp. 17–22. https://sites.google.com/site/nlp4cmc2015/program

[WAC-X/EmpiriST 2016] Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-26). http://aclweb.org/anthology/W/W16/W16-26.pdf.

Zinn, C. (2016). The CLARIN Language Resource Switchboard. In *Proceedings of the CLARIN Annual Conference*, *Aix-en-Provence, France*, https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf.