Introducing the CLARIN K(nowledge)-Centre for CMC and Social Media Corpora (CKCMC)

Egon W. Stemle[1], Jennifer-Carmen Frey[2], Alexander König[3], Achille Falaise[4], Tomaž Erjavec[5] and Harald Lüngen[6]

[1]*Institute for Applied Linguistics, Eurac Research, IT*
egon.stemle@eurac.edu
[2]*Institute for Applied Linguistics, Eurac Research, IT*
JenniferCarmen.Frey@eurac.edu
[3]*CLARIN ERIC, NL*
alex@clarin.eu
[4]*Laboratoire de Linguistique Formelle, FR*
afalaise@linguist.univ-paris-diderot.fr
[5]*Jožef Stefan Institute, SI*
tomaz.erjavec@ijs.si
[6]*Institut für Deutsche Sprache, DE*
luengen@ids-mannheim.de

The CLARIN Knowledge Centre for Computer-Mediated Communication and Social Media Corpora (CKCMC) offers expertise on language resources and technologies for Computer-Mediated Communication and Social Media. Its basic activities are to A) Give researchers, students, and other interested parties information about the available resources, technologies, and community activities, B) Support interested parties in producing, modifying or publishing relevant resources and technologies and C) Organise training activities. Additionally, the CKCMC manages the cmc-corpora.org website and helps in curating the CLARIN CMC Resource Family. The CKCMC can be reached via a Helpdesk (e-Mail). In this talk, we first want to introduce the CKCMC to the CMC-community; secondly, we would like to discuss with and get input from the CMC community on how to better serve our community goals or what further goals to pursue.

CLARIN Knowledge Centres (K-centres) are a cornerstone of the CLARIN Knowledge Infrastructure. They ensure the transfer of knowledge between all players involved in the construction, operation and use of the CLARIN infrastructure (Hinrichs & Krauwer, 2014). Their and thus the CKCMC's mission is to ensure that the available knowledge and expertise is made accessible in an organised way to both the CLARIN community and the social sciences and humanities research community more widely (https://www.clarin.eu/content/knowledge-centres).

The CKCMC is a distributed K-centre and supported by four institutions: The leading partner is the Institute for Applied Linguistics at Eurac Research, Italy (IAL); the other three partners are the Jožef Stefan Institute, Slovenia (IJS); the Laboratoire de Linguistique Formelle, France (LLF); and the Leibniz-Institut for the German Language, Germany (IDS). All institutes have been involved in CMC and Social Media corpus creation or curation projects, and they have long experience with computational linguistic processing, usage, and storage of standard and non-standard language data (Frey et al., 2016; Fišer et al., 2020; Chanier et al., 2014; Beißwenger et al., 2015). Together they cover most of the major language families in Europe and also have experience with low-resource and minority languages.

Frey et al. (2020) analysed the data management policies of 24 European CMC corpora according to the FAIR principles. The result showed that the prevalent data management policies were often only partly and almost never fully compliant with FAIR principles. In particular, when

creating a CMC corpus for the first time knowledge of (sometimes implicit) community standards may not have been available, which often led to non-interoperable or non-reusable data. Although the situation has improved, the CKCMC wants to strengthen efforts for depositing well-structured CMC corpora at (institutional) repositories for long-term research data preservation; and the CKCMC wants to facilitate community-driven efforts to raise awareness for all stages of FAIR research data management. This was showcased with the "Workshop: Data Management for FAIR CMC corpora" that was organised by the CKCMC at the CMC-Corpora Conference 2021. The CKCMC also wants to spread information about TEI-CMC (Beißwenger & Lüngen, 2020), a TEI customisation for the representation of CMC and Social Media corpora. Both goals foster interoperability between language resources as well as their analysis and automatic exploitation. Another interest is to observe the use of TEI-CMC in the community and the use of its expressive possibilities for the observed phenomena and the description of the data and metadata.

References

Beißwenger, M., Ehrhardt, E., Horbach, A., Lüngen, H., Steffen, D., & Storrer, A. (2015). Adding Value to CMC Corpora: CLARINification and Part-of-speech Annotation of the Dortmund Chat Corpus. 12–16. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4365

Beißwenger, M., & Lüngen, H. (2020). CMC-core: A schema for the representation of CMC corpora in TEI. *Corpus*, 20. https://doi.org/10.4000/corpus.4553

Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C. R., Hriba, L., Longhi, J., & Seddah, D. (2014). The CoMeRe corpus for French: Structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*, 29(2), 1. https://halshs.archives-ouvertes.fr/halshs-00953507

Fišer, D., Ljubešić, N., & Erjavec, T. (2020). The Janes project: Language resources and tools for Slovene user generated content. *Language Resources and Evaluation*, 54(1), 223–246. https://doi.org/10.1007/s10579-018-9425-z

Frey, J.-C., Glaznieks, A., & Stemle, E. W. (2016). The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts. In P. Basile, A. Corazza, F. Cutugno, S. Montemagni, M. Nissim, V. Patti, G. Semeraro, & R. Sprugnoli (Eds.), *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. https://doi.org/10.4000/books.aaccademia.1782

Frey, J.-C., König, A., Stemle, E., Falaise, A., Fišer, D., & Lüngen, H. (2020). The FAIR Index of CMC Corpora. In J. Longhi & C. Marinica (Eds.), *CMC Corpora through the prism of digital humanities*. L'Harmattan. https://api.zotero.org/users/332053/publications/items/R7Z5VHA2/file/view

Hinrichs, E., & Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 1525–1531. http://www.lrec-conf.org/proceedings/lrec2014/pdf/415_Paper.pdf