CLARIN Survey of CMC Resources and Tools

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

Abstract

With the growing volume and importance of computer-mediated communication, the need to understand its linguistic and social dimensions, along with CMC-robust language technologies is on the rise as well. This is reflected in the increasing number of conferences, projects and positions involving analysis of CMC in a wide range of disciplines in Digital Humanities, Social Sciences and Computer Science. As a result, a number of valuable CMC corpora, datasets and tools are being developed (Beißwenger et al., 2017) but unfortunately, due to non-negligible technical, legal and ethical obstacles, not many are being shared and reused.

Since it is the mission of CLARIN to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for researchers in Digital Humanities and Social Sciences (Krauwer and Hinrichs, 2014), it is our goal to have a good overview of the available resources and tools, to offer support to their developers to overcome the technical, legal and ethical obstacles and deposit them to the CLARIN infrastructure, as well as to the researchers with diverse backgrounds, such as linguistics, media studies, psychology etc., but also to interested parties from the educational, commercial, political, medical and legal sectors of the society who are interested in using them.

The first step in this direction was an interdisciplinary workshop[1] on the creation and use of social media which was organized within the Horizon 2020 CLARIN-PLUS project on 18 and 19 May 2017 in Kaunas, Lithuania. The aims of the workshop were to demonstrate the possibilities of social media resources and natural language processing tools for researchers with a diverse research background and an interest in empirical research of language and social practices in computer-mediated communication, to promote interdisciplinary cooperation possibilities, and to initiate a discussion on the various approaches to social media data collection and processing.

The workshop also served as a platform to conduct a survey[2] of corpora, datasets and tools of computer-mediated communication in the languages spoken in countries that are members and observers of CLARIN ERIC. Apart from identifying the existing resources and tools, our motivation was to establish to which extent they are accessible through the CLARIN infrastructure and how the information and accessibility of them could be further optimized from a user perspective.

In this talk, I will give an overview of the identified corpora, the smaller, more focused datasets and tools that are tailored to processing computer-mediated communication. The focus of the talk will be on the comprehensiveness of the provided metadata, level of availability and accessibility of the identified resources and tools and the degree of their actual or potential inclusion in the CLARIN infrastructure. I will also discuss the simple and long-term possibilities of enriching the current state of the infrastructure and provide guidelines for creating and depositing CMC resources with a CLARIN center.

Keywords: CLARIN ERIC, research infrastructure, language resources, NLP tools, computer-mediated communication

References

Beißwenger, M., Chanier, T., Erjavec, T., Fišer, D., Herold, A., Lubešić, N., Lüngen, H., Poudat, C., Stemle, E., Storrer, A., and Wigham, C. (2017). Closing a Gap in the Language Resources Landscape: Groundwork and Best Practices from Projects on Computer-mediated Communication in four European Countries. In *Selected Papers from the CLARIN Annual Conference 2016, Aix-en-Provence, 2628 October 2016, CLARIN Common Language Resources and Technology Infrastructure*, pages 1–18. Linköping University Electronic Press, Linköpings universitet.

Krauwer, S. and Hinrichs, E. (2014). The clarin research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531. European Language Resources Association (ELRA).

[1] https://www.clarin.eu/event/2017/clarin-plus-workshop-creation-and-use-social-media-resources

[2] https://office.clarin.eu/v/CE-2017-1064-Resources-for-computer-mediated-communication.docx