Introduction

The increasing popularity of Web 2.0 has resulted in an unprecedented surge of user-generated and social media content which is becoming a major source of knowledge and opinion, and is considered a catalyst of bottom-up communication practices that contribute towards the democratization of language. As a consequence, we are seeing a growing need for a thorough multidisciplinary understanding of this type of communication that is significantly shaped by the specific social and technical circumstances in which it is produced: rich in colloquialisms and foreign language elements, non-canonical spelling variants and syntax, idiosyncratic abbreviations and neologisms.

What is more, this form of highly participatory, interactive and multimodal communication is accompanied by easily accessible and rich (sociodemographic) data, which open a wide range of new and exciting research opportunities, not only in linguistics and natural language processing, but also in the digital humanities and social sciences, as well as bringing about new technical, linguistic and ethical challenges for scholars.

The major bottleneck in the dissemination of corpora of computer-mediated content is not a technical one, as text retrieval from user-generated and social media platforms, such as chats, forums, weblogs and tweets, on social network sites and in wikis, is generally straightforward and sometimes even facilitated by native APIs. Instead, the main reason for the low number of publicly available corpora is the unclear legal status of computer-mediated communication (CMC) data when distributed as a resource to the scientific community, which is further exacerbated by the rapidly changing terms of service by content providers.

To address these issues, a growing number of projects all over Europe have started to create CMC corpora which are intended to be made available to the scientific community, and thus close the "CMC gap" in the corpus landscape (Beißwenger et al. 2017). Since 2013, the annual conference series *CMC and Social Media Corpora for the Humanities*[1] has been dedicated to the discussion of best practices on all aspects of open issues regarding the development, annotation, processing and analysis of CMC corpora among researchers who are building and processing these, along with representatives of language resource infrastructure initiatives such as CLARIN and DARIAH, and researchers in linguistics, digital humanities and social sciences who are using CMC data and corpora for the analysis of CMC phenomena in different languages and for different genres. The results of previous conferences have been published in the form of a special issue of the *Journal of Language Technology and Computational Linguistics* (Beißwenger et al. 2014), a monograph *Corpus de communication médiée par les réseaux: construction, structuration, analyse* (Wigham and Ledegen 2017) and as online conference proceedings (Fišer and Beißwenger 2016).

[1] http://www.cmc-corpora.org/

For the first time, the call for papers for this monograph was open also to authors who did not present their work at the conference. It includes eight contributions that have been selected from a total of 16 submissions based on a double-blind peer review. They are written by 16 authors from 13 institutions in 13 different countries dealing with the creation of CMC corpora and with the analysis of CMC phenomena in 10 different languages. Five of them are original papers and three are extended papers from the 2016 edition of the CMC-Corpora Conference that was held in Ljubljana, Slovenia. They tackle a diverse range of research questions and use a rich set of approaches, which is why we have organized them into four broad thematic and methodological parts: lexical analysis of CMC, sociolinguistic analysis of CMC, conversation and conflict in CMC, and building and processing CMC resources.

Part 1: Lexical analysis of CMC

Maja Miličević, Nikola Ljubešič and Darja Fišer investigate the universalities and specificities of communication in social media environments in a comparative analysis of spelling conventions on Twitter for three closely related languages: Slovene, Croatian and Serbian. This corpus-based study reveals that words from closed classes tend to be more often realized in non-standard spellings than words from open classes; that character deletions are more frequent than insertions or replacements; and that tweets in the three focal languages deviate from the written standard norms to different degrees. The datasets created for the study can be used as resources for further investigation of non-standard spelling conventions in the three languages.

Mohamed Tristan Purvis compiles a WhatsApp dataset to analyse the vocabulary that Hausa-speaking chatters adopt when consciously referring to their chat environment. The author shows that the interlocutors represented in his dataset not only code-mix with common English terms, but also widely employ Hausa words adapted for specialized reference to the online environment. The study analyses lexical, semantic and sociolinguistic factors that promote or constrain the adoption and use of Hausa words in chat terminology.

Part 2: Sociolinguistic analysis of CMC

Lieke Verheijen addresses the power conflict between the overt prestige of the (written) standard language and the covert prestige of the language used among young CMC users. In order to determine how the language used by the Dutch youth in CMC differs from Standard Dutch, the author presents an extensive register analysis of about 400,000 tokens of digital texts, produced by 12–23 year-old adolescents and young adults in SMS, instant messages and tweets. The study focuses on the orthographic, typographic, syntactic and lexical features of such texts. The results offer linguistic profiles of Dutch written CMC language for four new media genres and two age groups.

Steven Coats investigates the extent to which English is used on Twitter in the Nordic countries, with a special focus on the link between gender and grammatical or part-of-speech frequencies, a link which has hitherto been considered mainly in the context of data collected from L1 Anglophone contexts. The study uses a corpus of English-language messages originating from the Nordic countries which has been built using the Twitter Streaming API. It applies automatic methods to disambiguate author gender, assign part-of-speech tags, and determines the relative frequencies of grammatical types by gender and country. The analysis shows that Nordic English-language discourse on Twitter diverges according to gender for a number of grammatical features. The analysis supports L1 findings pertaining to gendered differences in feature frequencies in English.

Part 3: Conversation and conflict in CMC

Tatjana Scheffler examines the linguistic and structural features of German Twitter conversations. The study reveals that many well-known dialog phenomena can also be observed on Twitter, while at the same time the writers avail themselves of more formal, written-like options, while some spoken-like features take on new meanings. An analysis of the dialog structure shows that Twitter is not a homogeneous conversational genre, but that different types of conversations must be distinguished. Overall, the paper outlines several perspectives for further research on Twitter conversations.

Lydia-Mai Ho-Dac, Veronika Laippala, Céline Poudat and Ludovic Tanguy analyse the linguistic features of conflicts which occur on Wikipedia talk pages where authors of Wikipedia articles coordinate the collaborative writing task and process. Using a large corpus of talk pages from the French Wikipedia, they try to determine the linguistic cues that may help to identify and characterize conflicts on talk pages with two methods: supervised automatic classification of conflicting vs. harmonious discussion threads and multidimensional analysis of the data, to highlight key features on the genre of Wikipedia talk pages at a global level. The results open up perspectives for future work on automatic classification and analysis of conversational phenomena in large CMC corpora.

Part 4: Building and processing CMC resources

Solange Aranha and Paola Leone discuss the creation of a special type of learner corpus that contains Voice-over-IP (VoIP) interactions in which an L2 learner and an expert in the target language meet on a weekly basis, and which are conducted partially in the learner's L1 and partially in the learner's L2 (Teledandem interactions). Research on the Teledandem system is growing rapidly, as it can help to better understand and foster various language learning processes. based on the example of the DOTI database, which is currently composed of 700 hours of video data from Teledandem sessions, the authors discuss the relevant metadata, especially the characteristics of the learning scenarios, the tasks and activities observed in these, and the CMC environment.

Michael Beißwenger, Tobias Horsmann and Torsten Zesch discuss options for improving the treatment of sparsely represented linguistic phenomena that are of special interest for the annotation of linguistic corpora. The authors present a case study in which they used a PoS tagger to find one particular phenomenon of that type, and discuss several approaches for improving the identification of occurrences of this phenomenon in chats and tweets. The case study is Based on a PoS-tagged data set of 230 instances of German verb-pronoun contractions which can be retrieved from the CLARIN repository at IDS Mannheim.

We hope that this book is as inspiring and enjoyable to read as it was to edit. Our work would not have been possible without the dedicated work of all the authors who submitted their contributions, and without the careful and insightful comments of the reviewers who operated under a very tight deadline: Špela Arhar Holdt, Adrien Barbaresi, Tomaž Erjavec, Axel Herold, Nikola Ljubešić, Nataša Logar, Julien Longhi, Harald Lüngen, Maja Miličvić, Céline Poudat, Müge Satar, Tatjana Scheffler, Egon W. Stemle and Ciara R. Wigham. We would also like to thank the language editor Paul Steed for polishing the manuscripts, and for all the support and good spirits provided by Matevž Rudolf and Jure Preglau from the Faculty of Arts Publishing House.

Darja Fišer and Michael Beißwenger
Ljubljana, Slovenia and Essen, Germany
31 July 2017

References

Beißwenger, Michael, Nelleke Oostdijk, Angelika Storrer and Henk van den Heuvel, 2014: Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. *Journal of Language Technology and Computational Linguistics* 2/2014. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf.

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

Wigham, Ciara R. and Gudrun Ledegen (eds.), 2017: *Corpus de Communication Médiée par les Réseaux. Construction, structuration, analyse*. Paris: L'Harmattan (Humanités numériques).