Reinhild Vandekerckhove*, Lisa Hilte, Darja Fišer and Walter Daelemans

Computer-mediated communication (CMC) and social media corpora: Introduction

https://doi.org/10.1515/eujal-2019-0032

This issue brings together language-centered studies on computer-mediated communication (CMC) and social media corpora.[1] They are illustrative of contemporary research interests in a very extensive research field that has strongly evolved over the past two decades: In the early days of CMC, much of the research was quite "anecdotal and speculative, rather than empirically grounded" (Herring 2004: 338) and in language-focused studies there was a predominant interest in the detection and description of prototypical features of the new genres (e.g. Crystal 2001). It took some time before the social and contextual embedding of these features and of CMC discourse in general was operationalized systematically as part of the research design. Androutsopoulos (2006: 430) discussed the reductive focus on the idiosyncrasies of the genre and argued that the time was ripe for "a user and community-centered approach, which is promising for a more complex theorizing of the social and contextual diversity of language use on the internet."

Since then the field has matured. CMC research has not only witnessed a boost, it got firmly embedded in linguistic disciplines like sociolinguistics, pragmatics, discourse analysis, and obviously computational linguistics. Furthermore, the

[1] Except for Hilte et al., the papers in this thematic issue represent a selection of the papers presented at the 6th Conference on Computer-Mediated Communication and Social Media Corpora, 17th-18th September 2018, University of Antwerp (Belgium). The annual conference series is dedicated to the collection, annotation, processing and exploitation of corpora of computer-mediated communication (CMC) and social media for research in the humanities.

***Corresponding author: Reinhild Vandekerckhove,** University of Antwerp, Department of Linguistics (CLiPS), Antwerpen, Belgium, E-Mail: reinhild.vandekerckhove@uantwerpen.be
Lisa Hilte, University of Antwerp, Department of Linguistics (CLiPS), Antwerpen, Belgium, E-Mail: lisa.hilte@uantwerpen.be
Prof. Darja Fišer, University of Ljubljana, Department of Translation, Ljubljan, Slovenia, E-Mail: Darja.Fiser@ff.uni-lj.si
Prof. Walter Daelemans, University of Antwerp, Department of Linguistics (CLiPS), Antwerpen, Belgium, E-Mail: walter.daelemans@uantwerpen.be

small data samples which much of the pioneering work relied on have been replaced by ever-growing corpora. The latter became an extra stimulus for computational linguistic and statistical data processing and contributed to the interdisciplinary character of the field. Even in language-focused CMC studies, this interdisciplinarity often extends far beyond the interaction of several linguistic disciplines (see e.g. Schwartz et al. (2013), which is also illustrative of the big data trend). While the present issue has a primarily linguistic focus, several papers reflect the interdisciplinary orientation of the field and the interest in substantial databases (see Coats; Flesch; Hilte et al., Longhi et al.). Yet contrary to the general scaling-up of the databases and a strong quantitative orientation in much of the present-day linguistic CMC and social media research, some of the contributions also show that fine-grained qualitative analyses of manually manageable datasets may still be a prerequisite for adequately capturing the subtleness or idiosyncracy of CMC pragmatics and the making of social meaning (e.g. Beißwenger & Pappert).

Apart from the linguistic and stylistic analyses of CMC-writing, content-oriented analyses are part of the field too, even within the language-centered approaches. Androutsopoulos (2011: 145) uses the term “digital networked writing” in order to emphasize “the dialogical and process-oriented character” of social media writing and argues that it derives its prototypical characteristics from several conditions, one of them being its “interpersonal and relationship-focused rather than subject-oriented” character. The interpersonal orientation is indeed predominant in many CMC genres and a strong determinant for both the media formats and discourse styles. However when tweeting or posting messages on discussion forums the participants tend to be strongly subject-focused as well. Therefore topic analysis (e.g. Schwartz et al. 2013) or “issue communication” (e.g. Praet et al. 2018) definitely is an issue in CMC research as well. The propagation or dissemination of themes and ideas via social media has received ample attention in CMC research, e.g. with topic modeling in computational linguistics. Again, this contributes to the interdisciplinarity of the field: quite a lot of research in this area finds itself at the intersection of political science, sociology and linguistics, as reflected in one of the contributions in this issue (see Longhi et al.)

As the above already suggests, the present issue reflects the diversity of the field in several respects. This is also the case with respect to the range of social media and CMC genres covered in the articles: some contributions focus on media for private and informal online interaction like WhatsApp (Hilte et al.; Lüngen & Herzberg) and Facebook Messenger (Hilte et al.), while others draw on media for more public communication like Twitter (Coats, Longhi et al., Lüngen & Herzberg), blogs (Lüngen & Herzberg) and discussion forums (Longhi et al.). Apart from these Lüngen & Herzberg also include Wikipedia talk pages for wiki-editors. The study by Flesch is based on a large corpus of comments posted by users of the social news site Reddit. Finally Beißwenger & Pappert operate in a much more controlled setting, i.e. a learning environment within a wiki platform.

The range of linguistic disciplines and methodological approaches in the present issue also deserves some more attention. First of all, some contributions have a distinct sociolinguistic focus: Flesch presents a quantitative sociolinguistic analysis of the social and ethnic correlates for the use of ‘tho’ (though) by Reddit users and examines its role in online identity construction. While she finds stronger effect sizes for the ethnicity of the users than for the other social correlates (with ‘tho’ being favored by Hispanics and Blacks), the appropriation of it by white young males presents an interesting case of social meaning making. Hilte, Vandekerckhove & Daelemans also present a sociolinguistic study that combines a quantitative and qualitative approach: they discuss the perception of social patterns in online writing by adolescents and confront perception with production: adolescents’ awareness of social patterns and their appreciation of particular social markers are compared to their (or their peers’) actual production of informal CMC. Strikingly, while the teenagers score high for age and gender detection, they manifest a very low awareness of stylistic differences related to educational background. Yet the latter variable appears to be a significant determinant of adolescent online writing. While Hilte et al. mainly focus on the social correlates and the social connotations of particular CMC-markers, Beißwenger & Pappert study the pragmatics of one of the most prototypical markers of the genre: emoji. In their qualitative analysis they discuss the use of emojis as part of politeness strategies and demonstrate how emojis contribute to facework. Their conclusion is that emojis operate on different levels of the organization of discourse to fulfil a range of functions on the semiotic, the pragmatic and the structuring level. Therefore they argue for a comprehensive pragmatic approach for the analysis of emojis in CMC interactions and for the systematic annotation of emoji functions in corpora.

The contribution of Coats has a strong linguistic focus – it incorporates morphological, phonological, semantic and spelling issues – , while at the same time laying bare the potential of computational linguistic methods for data extraction. It studies new and non-standard German verbal Anglicisms in a large Twitter corpus. The study introduces a method for automatic generation of an impressive amount of potential new verbal Anglicisms (in German) by using regular expressions. From a methodological perspective, this may set a standard for comparable research on lexical innovations and, in particular, the study of ‘new’ borrowings in other languages besides German. Moreover, just like the paper of Flesch, this paper shows the potential of big social media corpora for the study of linguistic innovation and more specifically for low-level phenomena.

Unlike the paper of Coats, the contribution of Longhi, Marinica & Després has no traditional linguistic focus at all, since it is much more content-oriented. It relies on textometry, or statistical text analysis, for detecting predominant lexical classes in the discourses of a political leader and his militant community. Both the temporality and the intensity of the use of particular terms are visualized so as to make clear potential interactions between the political discourse by the party leader himself and that of his militants (in one direction or the other) and the way political ideals might propagate. While the methodology is mainly computational linguistic, the study also relates to political science and communication sciences.

The final paper has a predominantly practical and methodological orientation: Lüngen & Herzberg present an analysis of different types of reply relations in CMC interactions and in doing so reveal the necessity of annotating these adequately so as to enable correct identification of relational structures between user contributions. The article does not only discuss the problem analysis, by focusing on the complexity of the reply relations and their formal realization, it also presents a way of dealing with these issues in the form of a proposal for annotating all of the relevant parameters of these reply relations within the TEI annotation framework. Consequently, the paper has a strong technical focus.

The issue closes with a short project description by Beißwenger, Fladrich, Imo and Ziegler: they report on findings from the *MoCoDa2* project, which focuses on the collection of private CMC-data and – interestingly – involves CMC-users not only as donators but also as editors of their data.

When highlighting the diversity of the field and the way this is reflected in the present issue of EuJAL, it should also be mentioned that the issue contains contributions on several languages. Several papers are based on German data (Coats; Beißwenger & Pappert; Lüngen & Herzberg), though not exclusively from Germany: Coats systematically includes Austrian and Swiss data as well. Longhi et al. deal with French data. Hilte et al. study the language perceptions and language behaviour of adolescents whose native language is Flemish Dutch. And finally the contribution of Flesch on the use of 'tho' relies on English data that were mainly produced by Americans with diverse ethnic backgrounds.

It may be clear from the above that CMC offers a very dynamic and challenging research field for researchers with an applied linguistic orientation: social media discourse is the locus for linguistic innovation, for identity construction, for social meaning making and facework, and for the exchange and dissemination of ideas and ideologies. The present issue covers all of these dimensions. It also presents practical and methodological opportunities and challenges. While significant progress has been made in the past two decades, the field could advance further by tackling methodological issues like e.g. the need of adequate annotation of relational structures between participants. Therefore the study of CMC and social media corpora remains an invitation for interdisciplinary cooperation.

References

Androutsopoulos, Jannis. 2006. Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics* 10. 419–438.

Androutsopoulos, Jannis. 2011. Language change and digital media: A review of conceptions and evidence. In Tore Kristiansen & Nikolas Coupland (eds), *Standard languages and language standards in a changing Europe*, 145–161. Oslo: Novus Press

Crystal, David. 2001. *Language and the Internet*. Cambridge: Cambridge University Press.

Herring, Susan. 2004. Computer-mediated Discourse Analysis. In Bara, Sasha, Kling, Rob & Gray, James H. (eds.), *Designing for Virtual Communities in the Service of Learning*, 338–376. New York: Cambridge University Press.

Praet, Stiene, Walter Daelemans, Tim Kreutz, Peter Van Aelst, Stefaan Walgrave & David Martens. 2018. Issue communication by political parties on Twitter. *Data Science, Journalism and Media* (Association for Computing Machinery). 1–8.

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, & Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* 8(9). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791 (accessed 3 Juin 2019).