RiCL Research in Corpus Linguistics

Computer-mediated communication (CMC) and social media corpora[1]

Mario Cal Varela – Francisco Javier Fernández Polo – Ignacio M. Palacios Martínez
University of Santiago de Compostela / Spain

Abstract – The study of computer-mediated communication (CMC) has received extensive attention in recent years, due to its impact on human communication and the immediacy of its form. This introduction briefly reports on some of the changes that CMC has undergone lately. The focus is on those topics currently considered to be central to the field, such as questions of identity and ideology, (im)politeness and face, humour, group creation and affiliation, verbal violence, cyberbullying, etc. Some observations are also made on the challenges that the compilation of CMC corpora poses for linguists, ranging from data copyright, anonymisation and representativeness to distinctive features of CMC texts, namely multimodality, non-standard language and non-sequential organisation. It also introduces each of the eight papers selected for this special issue of *Research in Corpus Linguistics*, highlighting their specific contribution to the field of CMC studies.

Keywords – computer-mediated communication, digital communication, digital genres, social media, online interaction

Computer-mediated communication (henceforth, CMC) can be roughly defined as human communication through the new technologies. The study of CMC is a highly interdisciplinary field borrowing concepts and methods from linguistics, sociology or computer science, among others. Specialists have given the field other more encompassing names —for instance, 'technology-mediated communication' (Dynel and

[1] For generous financial support, we are grateful to the following institutions: The *University of Santiago de Compostela*, The *Spanish Ministry of Science and Innovation* (grant PID2021-122267NB-I00), the *European Regional Development Fund* (grant PID2021-122267NB-I00) and the *Regional Government of Galicia* (Consellería de Educación, Cultura e Universidade, grant ED431B 2021/02). We would also like to warmly thank the reviewers of the papers of this Special Issue for their insightful comments and suggestions for improvement of the original manuscripts: Isabel Balteiro Fernández (University of Alicante), Marie-Louise Brunner (Trier University of Applied Sciences), Marta Carretero Lapeyre (Complutense University of Madrid), Turo Hiltunen (University of Helsinki), Sven Leuckert (TU Dresden), Paula López Rúa (University of Santiago de Compostela), Carmen Maíz Arévalo (Complutense University of Madrid), Pedro Martín Martín (University of La Laguna), Pilar Mur Dueñas (University of Zaragoza), Paloma Núñez Pertejo (University of Santiago de Compostela), Mercedes Querol-Julián (International University of La Rioja), José Sánchez Fajardo (University of Alicante), Laurel Stvan (University of Texas at Arlington), Crispin Thurlow (University of Bern) and Eva Triebl (University of Viena). Words of recognition and appreciation also for the general editors of *Research in Corpus Linguistics*, Carlos Prado Alonso and Paula Rodríguez Puente, for making things easy and supervising the whole process so efficiently.

Chovanec 2015), 'online communication' (Collins 2019), or 'digital communication' (Zappavigna 2012; Garcés-Conejos Blitvich and Bou-Franch 2019)— which, it is claimed, reflect best the wide variety of technologies, media, and highly multimodal nature of present-day mobile communication technology. However, CMC remains a popular umbrella term (Zappavigna 2012), frequently found in monographs (Herring *et al.* 2013), book series, reference works (e.g., *Wikipedia*) and specialised journals and conferences.

As a research area, CMC has undergone significant changes in view, first (and naturally), of the evolution of the technologies themselves, but also of the new interests and research paradigms, particularly of linguistics. Methodologically, research on CMC has traditionally favoured qualitative methods, including discourse analysis, multimodal analysis, critical discourse analysis, conversation analysis and others (Sung *et al.* 2021). This bias may result from "the restrictions that social media put on a quantitative approach", as a specialist recently complained (personal communication), but may also be explained by the socio-pragmatic agenda that has become popular in CMC since the early 2000's (Herring *et al.* 2013). The very name of the field —'computer mediated discourse analysis' (Herring 2004), 'new media sociolinguistics' (Thurlow and Mroczek 2011), etc.— reflects the theoretical, methodological and thematic preferences of the authors.

Early interest in the characteristic features of CMC (such as expressive uses of punctuation and emoticons, pragmatic rules of turn-taking, discourse organisation, etc.) has been expanded and approached from a socio-pragmatic perspective, in recognition of the fact that "digital texts are grounded in situated social and cultural practices" (Johansson *et al.* 2021: 3). A major strand of research refers to how participants engage in interaction and how forms of participation reflect aspects of the communicative situation, including personal identity (age, gender, origin, etc.), participant role or social status, but also the specific technological constraints, as well as broader issues of ideology and social power. Popular topics in CMC monographs and journals include issues on (im)politeness and face, humour, group-creation or affiliation, creativity or innovation, but also cyberbullying, trolling, verbal violence, or disinformation (Rüdiger and Dayter 2020).

Issues of identity and ideology are particularly cherished. Research on social media has shown special interest in the way that we "construct who we are and how we relate

to others" (Garcés-Conejos Blitvich and Bou-Franch 2019: 10), and how existing ideologies shape and are shaped by our communicative practices.

Ethical questions, in general, are at the core of research in CMC. For a start, it is difficult to establish a clear boundary between what is public and private in these contexts (Garcés-Conejos Blitvich and Bou-Franch 2019). Questions of participant consent and anonymisation have preoccupied CMC corpora compilators from the start (Beißwenger and Storrer 2008). Major ethical questions are still central in today's CMC research agendas. Issues of security and deception have always plagued digital communication. More importantly, critical and ethical approaches are justified by the huge potential of social media to exert manipulation and control on its users, and some have argued for a focus on the study of language in use, trying to illuminate social and cultural problems and inequalities (Thurlow and Mroczek 2011).

Quantitative methods include corpus linguistics (Beißwenger and Storrer 2008; Baker 2009; Sun *et al.* 2021). Quantification and corpora naturally play a key instrumental role in the analysis and substantiation of claims in qualitative studies. Corpus-based approaches have been around from the start, in studies comparing digital and non-digital communication (Biber and Conrad 2009), or describing the characteristic features of specific digital genres (Zappavigna 2012).

The compilation of CMC corpora poses new and significant challenges (Collins 2019), ranging from traditional issues of copyright, anonymisation, or representativeness (Laitinen and Lundberg 2020), to issues related to the special nature of CMC texts: non-standard language, complex multimodality, non-sequential organisation, or the uncertain nature of participants are some of the complicating factors in CMC corpora compilation, requiring new solutions (Beißwenger and Lüngen 2020). Although the internet is an immense source of linguistic data, paradoxically access to quality data for a carefully constructed corpus remains a perennial problem. Recent restrictions on the access to *Twitter/X* data clearly do not help.

Some of the issues and topics above are discussed in this special issue, in which we present a sample of state-of-the-art research on CMC corpora, intended to showcase some of the new trends in this vast research field. Many of the contributions were originally presented at a special conference on CMC corpora celebrated at the University of Santiago de Compostela in September 2022. As a follow-up to the conference, a special call was first issued for participants to submit an elaborated version of their research for

a special issue on the topic, which was then extended to other specialists who had not participated in the event.

All the articles present corpus-based empirical research into CMC and social media corpora, representing a wide variety of topics, media and communicative contexts, approached from diverse theoretical perspectives, including sociolinguistics, discourse analysis, pragmatics or genre analysis. The various articles, mostly on English usage online by both native and non-native speakers, provide a good illustration of the multidisciplinary and methodologically innovative nature of CMC research (Coats; Verheijen and Mauro). They also demonstrate how the analysis of CMC corpora may shed new light on classic topics in Linguistics, like lexical creativity and innovation (Spina *et al.*), syntax (Doval-Suárez and González-Álvarez), or language variation (Romasanta), while furnishing a clear illustration of the deep social engagement that characterises the field (Foley; Loureiro-Porto and Ariza-Fernández; Villares Maldonado).

This special issue starts with three very interesting **contributions on users' experiences of social media**.

In his paper, **Steven Coats** uses cutting-edge natural language processing tools to look at public online interaction with local governments from the perspective of computational social science. He applies computational techniques to analyse a huge sample of over 20,000 video transcripts and over 190,000 public comments on those videos drawn from the *Corpus of North American Spoken English* (CoNASE; Coats 2023), a 1.3-billion-word corpus of transcripts of videos uploaded to the YouTube channels of municipalities and other local government entities in the US and Canada. He shows how transformer model-based tools such as summarisation of discourse, topic modelling and sentiment analysis can be used meaningfully to analyse public reactions to online content and provide useful information to, for example, guide local governments in their public communication policies in order to increase civic engagement.

Jennifer Foley reports on a pilot study of a 20,000-word specialised corpus of blog posts in which she explores how users resort to metaphorical expressions to conceptualise social media and its effect on mental health and wellbeing. She shows that while conventional metaphors often provide a negative evaluation of social media, they may also be used to highlight potential benefits. All in all, she demonstrates that the analysis of metaphors, in combination with approaches from fields studying people's thought

processes and emotions, may prove a valuable tool to investigate how social media is used to deal with mental illness and to identify both benefits and risks.

Verheijen and Mauro's paper represents a novel contribution on one of the most popular topics in CMC, emojis. They investigate emoji literacy and use in children compared to adults, additionally comparing the effect of a number of variables —age, gender and smartphone ownership— on the number, position and meaning of emojis for this specific age-group. To investigate the topic, they use a very innovative experimental method to collect their data, where participants are asked to add emoji magnets to a series of social media messages printed on a board. While children's use of emoji, in general, is similar to that of adults, the study reveals interesting differences, not only in their use but also in their interpretation across different groups.

The next two articles focus **on participants' management of interaction in CMC**, or the kind of communicative strategies they use to enhance interpersonal relations, which are central to the functioning of virtual communities.

Villares Maldonado explores an emergent digital genre, the *Twitter* conference presentation (TCP), showing how digital communication is changing the communication practices of specialised discourse communities. Her analysis focuses on the discussion section (TCDS) following the TCP itself. She combines a quantitative and qualitative approach, to shed light on the vast amount of interactional work that is realised by participants to preserve interpersonal relationships in this type of event. While discussion sessions in *Twitter* conferences basically share organisation and purpose with discussions in presential conferences, TCDS participants use both digital and *Twitter*-specific affordances to fulfil major functions of the genre —knowledge construction, community building and self-promotion— and compensate for the limitations of the medium.

Doval-Suárez and González-Álvarez analyse 165 instances of concessive clauses headed by *but* drawn from the *Santiago University Corpus of Discussions in Academic Contexts* (SUNCODAC 2021), a collection of student online discussions in which participants provide critical feedback to their peers. The authors show that these structures can occur in a diversity of interactive/semantic patterns, and also that they play an important role, in combination with other politeness strategies, in collaborative pedagogical contexts. Their detailed analysis of the co-occurrence of these structures with hedges, boosters, positive and negative sentiment words and pronominal forms reveals slight differences in interaction style which may be related to gender, and shows that

concessives are an interesting feature to focus on when tracking changes in the dynamics of learning communities over time.

The last three papers present research on various **key issues in CMC sociolinguistics: language change, language and gender, and geographical variation**.

Spina *et al.*'s paper is concerned with lexical change and innovation in contemporary Italian micro-blogging by using a large sample of geotapped tweets from the 2002 Italian *Twitter* timeline. More than 700 tokens are identified in the analysis as possible neologisms which are then classified under 14 different groups of lexical creation that cover a wide range of word-formation processes from suffixation, univerbation, transcategorisation to acronymic derivation, redefinition and tmesis. Out of all these, orthographic variation, suffixation, loanwords and blends are the most frequent resources that Italian uses for lexical creation. In light of the data obtained, the authors come to the conclusion that lexical creativity and innovation, amusement and attention-seeking seem to be the prevailing criteria in the coinage of these items rather than the real need of defining and identifying new concepts, events, or situations. In fact, the majority of these terms serve to convey discursive functions such as irony, intensification and emphasis.

In their paper **Loureiro-Porto and Ariza-Fernández** evince how *X* profiles can be regarded as valuable tools for the study and understanding of linguistic patterns connected with social trends, gender equality and network relations being two cases in point. To this aim, they investigate the usage of non-binary pronouns such as generic *they*, rolling pronouns *they/she* and neopronouns (ZE or XE) within the non-binary community by closely examining a sample of 6,432 *X* bios extracted with the analytic platform *Followermonk*,[2] which provides information about *X* users, their followers, social authority and various other metrics. The results show that, contrary to what could be expected, no major divergences in the use of these non-binary pronouns are identified across different US regions despite important ideological differences. The use of rolling pronouns seems to be the preferred option while neo-pronouns and monopronoun usage (e.g. *they*) are rare. Moreover, single pronouns tend to be accompanied by their accusative form in contrast to rolling pronoun users who tend to opt for the opposite trend.

Finally, **Romasanta** focuses on non-categorical syntactic variation in internet language by closely analysing data from blogs, websites, forums and comments as part

[2] https://followerwonk.com/

of the *Corpus of Global Web-Based English* (GloWbE; Davies 2013). For this purpose, she studies how the geographical area of internet users of several English varieties such as Indian English, Singaporean English, Sri Lankan, Bangladeshi, Malaysian, Philippine, Pakistani, British and American English may affect the use of the clausal complementation patterns available for the verb *regret* as regards the variation between finite *that*-clauses and nonfinite *-ing* clauses (*you will regret that you went to Lahore vs. you will regret going to Lahore*). The analysis of a sample of over 10,000 tokens shows that the geographical origin factor has a clear impact on the complementation system of this verb, regarding the variables that condition variability and the preferences for particular patterns. This means that geographical distance between the different varieties conditions the similarities or differences among the varieties considered thus permitting making a distinction between three main geographical areas: 1) South Asia including India, Sri Lanka, Pakistan and Bangladesh, 2) South-East Asia with Singapore, Malaysia and the Philippines, and 3) East Asia (Hong Kong).

We believe that the wide variety of topics and the interesting results presented in this collection of studies will be of special interest to those specialists in CMC, as well as to those readers who would like to initiate their research in this fascinating area of communication and linguistic studies.

REFERENCES

Baker, Paul ed. 2009. *Contemporary Corpus Linguistics*. London: Continuum.

Beißwenger, Michael and Harald Lüngen. 2020. CMC-core: A schema for the representation of CMC corpora in TEI. *Corpus* 20. https://doi.org/10.4000/corpus.4553

Beißwenger, Michael and Angelika Storrer. 2008. Corpora of computer-mediated communication. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics. An International Handbook.* Berlin: Mouton de Gruyter, 292–308.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge: Cambridge University Press.

Coats, Steven. 2023. Dialect corpora from *YouTube*. In Beatrix Busse, Nina Dumrukcic and Ingo Kleiber eds. *Language and Linguistics in a Complex World*. Berlin: De Gruyter, 79–102.

Davies, Mark. 2013. *Corpus of Global Web-based English: 1.9 Billion Words from Speaker in 20 Countries (GloWbE)*. https://www.english-corpora.org/glowbe/

Collins, Luke. 2019. *Corpus Linguistics for Online Communication: A Guide for Research*. London: Routledge.

Dynel, Marta and Jan Chovanec. 2015. *Participation in Public and Social Media Interactions*. Amsterdam: John Benjamins.

Garcés-Conejos Blitvich, Pilar and Patricia Bou-Franch. 2019. Introduction to analyzing digital discourse: New insights and future directions. In Patricia Bou-Franch and Pilar Garcés-Conejos Blitvich eds. *Analyzing Digital Discourse*. Cham: Springer, 3–22.

Herring, Susan C. 2004. Computer-mediated discourse analysis: An approach to researching online communities. In Sasha A. Barab, Rob Kling and James H. Gray eds. *Designing for Virtual Communities in the Service of Learning*. Cambridge: Cambridge University Press, 338–376.

Herring, Susan C., Dieter Stein and Tuija Virtanen eds. 2013. *Pragmatics of Computer-Mediated Communication*. Berlin: Mouton de Gruyter.

Johansson, Marjut, Sanna-Kaisa Tanskanen and Jan Chovanec. 2021. Practices of convergence and controversy in digital discourses. In Marjut Johansson, Sanna-Kaisa Tanskanen and Jan Chovanec eds. *Analyzing Digital discourses: Between convergence and controversy.* Cham: Springer, 1–24.

Laitinen, Mikko and Jonas Lundberg. 2020. ELF, language change and social networks: Evidence from real-time social media data. In Anna Mauranen and Svetlana Vetchinnikova eds. *Language Change: The Impact of English as a Lingua Franca*. Cambridge: Cambridge University Press, 179–204.

Rüdiger, Sofia and Daria Dayter. 2020. The expanding landscape of corpus-based studies of social media language. In Sofia Rüdiger and Daria Dayter eds. *Corpus Approaches in Social Media Studies in Corpus Linguistics*. Amsterdam: John Benjamins, 1–12.

Sun, Ya, Gongyuan Wang and Haiying Feng. 2021. Linguistic studies on social media: A bibliometric analysis. *SAGE Open* 11/3:1–12.

SUNCODAC. 2021. *Santiago University Corpus of Discussions in Academic Contexts*. Santiago de Compostela: University of Santiago de Compostela. http://www.suncodac.com

Thurlow, Crispin and Kristine Mroczek. 2011. *Digital Discourse: Language in the New Media*. Oxford: Oxford University Press.

Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Bloomsbury.

Corresponding author
Ignacio M. Palacios Martínez
University of Santiago de Compostela
Department of English and German Philology
Avenida de Castelao, s/n
15872 Santiago de Compostela
Spain
E-mail: ignacio.palacios@usc.es