RiCL

Research in Corpus Linguistics

RiCL 13/1 (2025)

Special Issue

"Computer-mediated communication (CMC) and social media corpora"

edited by Mario Cal Varela, Francisco Javier Fernández Polo and Ignacio M. Palacios Martínez

aelinco
Asociación Española de Lingüística de Corpus

RiCL 13/1 (2025)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

13/1 (2025)

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

RiCL Research in Corpus Linguistics

Commenting on local politics: An analysis of *YouTube* video comments for local government videos

Steven Coats
University of Oulu / Finland

Abstract – This study compares the content of transcripts of videos uploaded by local governments with the comments on those videos, utilizing three transformer-model-based techniques: summarization of the discourse content of video transcripts, topic modeling of summarized transcripts, and sentiment analysis of transcripts and of comments. The analysis shows that some types of video content, for example those dealing with music or education, are more likely to attract positive comments than content related to policing or government meetings. In addition to their potential relevance for local government outreach, the study may represent a viable exploratory method for comparison of online video content and written comments in the context of computational social science analyses of user interaction and commenting behavior.

Keywords – *YouTube*; comments; ASR transcripts; local government; transformer models; topic modeling; sentiment analysis

1. Introduction and background[1]

Comparison of the discourse content of video streams with comments on those streams represents an under-researched topic in studies of Computer-Mediated Communication (henceforth, CMC) and discourse. Over the last two decades, there has been a noticeable transition towards a greater reliance on CMC environments, a shift encompassing various forms of communicative interactions and interactive registers. Notably, civic engagement at the local community level is increasingly conducted online, a tendency facilitated by increased access to the communicative affordances of online platforms and accelerated in the early 2020s by the Covid-19 pandemic. While feedback via CMC provides citizens with a means to express their satisfaction and their concerns about the workings of local government and issues of local importance, online commenting differs from traditional forms of citizen engagement. Comments on video streams or recordings exhibit

[1] The author would like to extend thanks to Finland's *Centre for Scientific Computing* for providing computational resources, and to two anonymous reviewers for their helpful comments.

communicative features that reflect the interactive parameters of the medium as well as aspects of the online userbase in ways that make them difficult to compare with traditional feedback forms. Nevertheless, public comments are important sources of information for local governments and other organizations, and gauging public sentiment towards local government ordinances, initiatives, services, and news/information is an important aspect of responsible and successful governance.

YouTube comments have attracted substantial research attention and, in recent years, their linguistic and interactive properties have been the subject of qualitative, quantitative, and corpus-based analysis from a variety of theoretical and methodological perspectives. Studies of *YouTube* comments have investigated questions of commentator stance and addressee (e.g., Bou-Franch *et al.* 2012; Dynel 2014; Herring and Chae 2021), discourse pragmatic concerns such as impoliteness and flaming (e.g., Andersson 2021; Lehti *et al.* 2016), or the relationship between video content, popularity, and commenting behavior (e.g., Siersdorfer *et al.* 2014; Ksiazek *et al.* 2016), among other topics. However, despite the diversity of approaches, few studies have compared comments specifically with the language content of videos, and *YouTube* channels of governmental organizations have not been a primary focus.

For this study automatic speech recognition (henceforth, ASR) transcripts from the *Corpus of North American Spoken English* (CoNASE; Coats 2023), were assessed in terms of sentiment, and summarized using a transformer model. The summarized transcripts were then assigned to topics using BERTopic (Grootendorst 2022), a suite of topic modeling scripts that utilizes large, context-sensitive transformer models. All available comments for the corresponding videos were then retrieved and assessed in terms of sentiment using the same model as used for transcripts, namely, twitter-roberta-base-sentiment-latest (Camacho-Collados *et al.* 2022), a fine-tuned version of RoBERTa-large (Liu *et al.* 2019). The study represents an exploratory approach to the following research questions:

1) What are the main topical concerns of local government meetings in North America?
2) What is the relationship between topic and the discourse content of transcripts in terms of sentiment?
3) What is the relationship between topic and the discourse content of comments in terms of sentiment?

The study shows that certain video topics, as determined by the summarization-topic modeling procedure, are more likely to represent positive sentiment as well as to attract positive comments. The analysis demonstrates how transformer models can be applied to publicly accessible data in order to assess trends and attitudes in the public sphere, and as such represents a method that may be relevant not only for the study of local governance, but for investigations of many types of online interaction. In terms of civic engagement, the results may help policymakers direct their social media outreach efforts towards the creation of content that is more likely to elicit viewer responses such as commenting and liking. Because communities with engaged citizens are more likely to exhibit positive traits such as increased government accountability or societal inclusiveness (Gaventa and Barrett 2012), engagement represents a desideratum of local government policymakers.

In a broader perspective, the comparison of the discourse content of videos and streams with comment content is relevant for the empirical study of discourse in terms of multimodal communicative pragmatics. The study exemplifies the use of corpus data and transformer models for social research and demonstrates an analytical approach for understanding the relationship between comments and video content. As such, it also represents an example of linguistic data science research at the intersection between language studies, social science, quantitative data analysis, and corpus-based computational sociolinguistics (Schmid 2020; Grieve *et al.* 2023; Coats and Laippala, 2024).

The article is organized as follows: Section 2 discusses some previous research into commenting behavior and comments on *YouTube* videos. Section 3 describes the data used in the study and the methods used to gauge sentiment in the transcripts and comments and assign transcripts to topics. Section 4 presents the largest topics in the transcript material and compares sentiment scores in the transcript material with sentiment scores in comments. In Section 5, the results are discussed and interpreted and several caveats pertaining to the data, methods, and interpretations are noted.

2. Previous research

Commenting behavior in CMC has been extensively studied. Early research investigated communicative and linguistic aspects of comment threads on bulletin boards and as responses to edited texts, such as news articles. More recent studies, a few of which are discussed below, have focused on commenting behavior on image-, video-, or live stream-hosting platforms such as *Instagram*, *YouTube*, *Twitch*, *TikTok*, and others.

2.1. Interactivity, pragmatics, and modeling of comments

Classifications of comments in terms of addressivity patterns and pragmatic functions have been the focus of linguistic studies of commenting behavior, for example, utilizing the theoretical and methodological framework of Conversation Analysis (Bou-Franch *et al.* 2012). Analyses of comment structure and content can be complicated by the, at times, unclear addressivity patterns within comment threads: that is, individual comments can be directed towards page content in general, towards individuals identified in the content on a page, towards other commentors on the page in general, or towards specific users/commentors, among other configurations. A basic distinction can be drawn between comments which are directed to the main content of a page such as the video or news text it presents and comments directed towards other comments, a distinction for which Ksiazek *et al.* (2016) suggest the terms 'user–content interactivity' and 'user–user interactivity'. In addition to different addressivity configurations, comments for some online platforms often make use of emoticons, emoji, and animated graphicons whose semantic and pragmatic values are not always easy to analyze. Herring and Dainas (2017), for example, analyzed the use of graphicons such as emoji and reaction image gifs in a corpus of *Facebook* posts, classifying them into five pragmatic categories. 'Reaction' usages, in which a graphicon is used in a stand-alone manner without accompanying text, were most common, followed by 'tone' usages, in which the images could be interpreted to be modifying the text content of the post.

Predictive modeling has been used to interpret patterns of comments. Häring *et al.* (2018), for example, created two large corpora of German-language comments on news articles and used word embeddings to train a classifier to distinguish between 'non-meta' and 'meta' comments (i.e., comments which address the content of the news article and comments which are directed towards the article author, the publisher, readers on the

news platform, the moderator of the comment space, or others). Ksiazek (2018) analyzed 330,000 comments on almost 2,000 news articles about diverse topics from English-language news websites. After articles were categorized into 25 different topics, content word frequencies from *Linguistic Inquiry and Word Count* (LIWC),[2] a tool which assesses the linguistic and psychological dimensions of text based on aggregate content word counts (Tausczik and Pennebaker 2010), were used to measure the civility and hostility of comments. Using a hierarchical regression model, he found that some news topics, such as the Tea Party, healthcare, and government budgets, were more likely to generate larger numbers of comments overall, whereas others, such as gun control or foreign policy, were more likely to attract negative or hostile comments. Krohn and Weniger (2019) created a model to predict the size of comment threads based on data from *Reddit*, a platform in which much of the content consists of hierarchically arranged user comments. Their model, which included post title, author, and other properties of seed posts, predicted the size and temporal dynamics of comment threads; they report improved results compared to baseline models.

2.2. YouTube comments

Commenting on *YouTube*, which as a platform has been characterized as a kind of mediated quasi-interaction (Bou-Franch *et al.* 2012), can occur with a variety of addressivity configurations (Dynel 2014; Herring and Chae 2021). As of 2023, *YouTube* comment threads have a maximum depth of two. Top-level comments are shown in order of recency or popularity directly under a video; replies to comments are shown indented under top-level comments (see Figure 1).[3]

[2] https://www.liwc.app/

[3] Please note that Figure 1 does not represent a real video or real comments but was created for illustrative purposes.

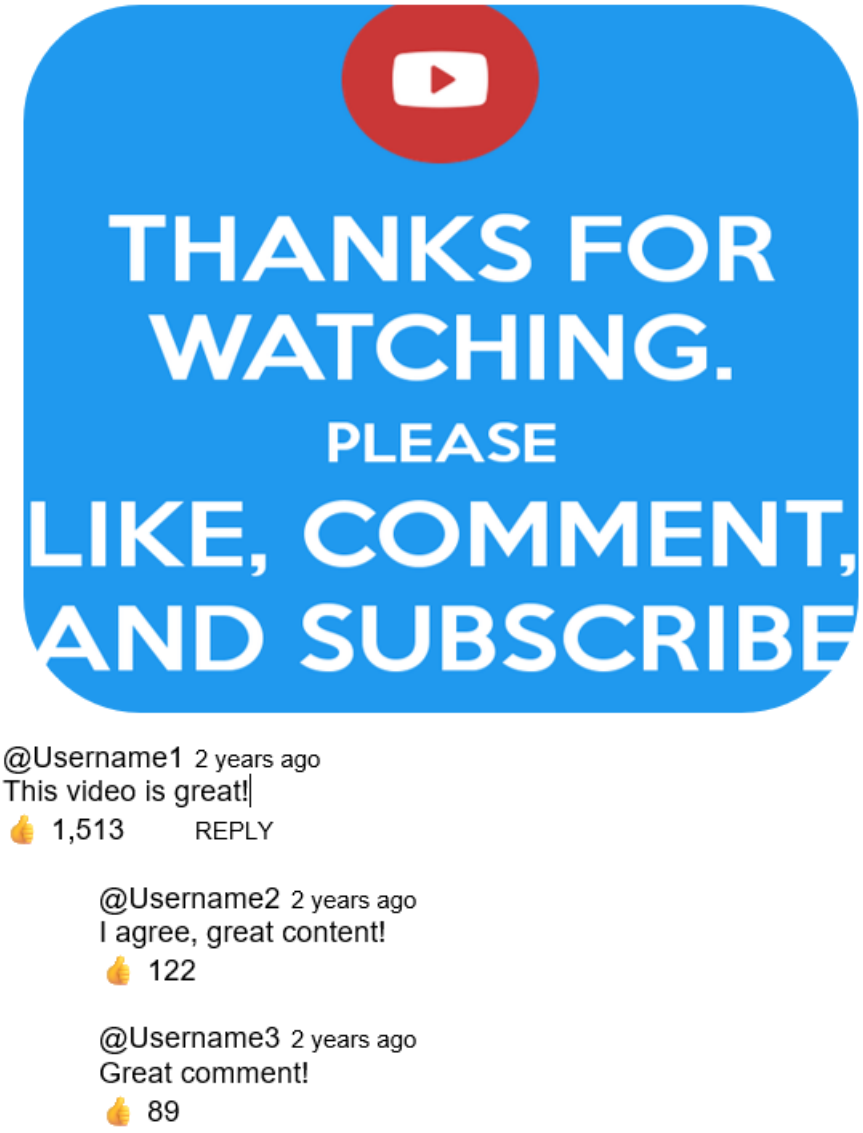

Figure 1: Schematic representation of a *YouTube* video and comment structure

As is the case with studies of other comment-based CMC, analyses of *YouTube* commenting behavior have been undertaken from qualitative and descriptive perspectives, as well as by building predictive models.

Qualitative studies include Goode *et al.* (2011), who analyzed 30 videos in the *YouTube* channels of eight mainstream news outlets. Investigating whether *YouTube* comments on news videos could represent an idealized Habermasian "public sphere" that enables positive civic participation and dialogue, they found that, on the contrary, *YouTube* comment sections tend to be an "unruly" place, characterized by expressions of anger, boredom, or vulgarity, with a "low signal-to-noise ratio" (Goode *et al.* 2011: 611). Bou-Franch *et al.* (2012) hand-coded 300 comments on two *YouTube* videos, comprising almost 12,000 words in total, for a variety of turn-maintenance devices described in

previous CMC research or derived from concepts developed in Conversation Analysis (essentially, whether a comment refers to the immediately preceding comment, to some other comment, or to the video on the page). They classified most comments as "adjacency turns" (Bou *et al*. 2012: 502) which referred to the immediately preceding comment. Lehti *et al.* (2016) described types of impoliteness in the comment thread of a well-known *YouTube* video from 2014. Herring and Chae (2021) discussed addressivity in comment threads on *YouTube*, noting that it is not always obvious to whom a comment is directed. Qualitatively analyzing 200 comments for each of three *YouTube* videos, they found that the largest proportion of comments are free-floating, without a specific addressee. Comments can be directed to speakers in the video, to other commenters on *YouTube*, to the *YouTube* platform, or to speakers in embedded videos, for video clips that include embedded content. Similarly, Cotgrove (2022) compiled a corpus of 3m *YouTube* comments from German-language youth-oriented videos to analyze lexical, grammatical, and discourse features of online youth language.

Quantitative and predictive modeling approaches have also been employed for the study of *YouTube* comments. In Schultes *et al.* (2013), comments for a pseudo-random sample of 304 *YouTube* videos were assigned class labels ('discussion post', i.e., a post containing content directed at another comment/user; 'inferior comment', containing insults, offensive statements, or short, emotional replies; or 'substantial comment', non-offensive comments directed towards the video's content) on the basis of features such as comment length in number of tokens, presence of offensive or emotional words or of emoticons, lexical overlap with the title of the video, and other features. They found that labels generated in this manner could be used to train a classifier to achieve high internal consistency when predicting comment type. In addition, they considered the relationship between these labels and the like/dislike ratio of videos. Discussion post comments were found to be the strongest predictor of likes, while inferior comments were found to better predict dislikes. It should be remarked, however, that *YouTube* comments at the beginning of the 2010s were a wilder place than at the present time, with relatively unsophisticated automatic filters on the platform making it possible to post a wider variety of potentially objectionable content (see, e.g., Nycyk 2012, who provides examples of abusive comments that are no longer encountered on the platform). Siersdorfer *et al.* (2014) considered comments on *YouTube* videos and on *Yahoo News* articles in terms of their aggregate ratings (like/dislike ratios) and how these corresponded to comment textual

content. They found that comments with higher ratings tended to include positive terms such as *love*, *greatest*, or *perfect*, whereas those with low ratings included negative terms such as *retard* or *idiot*.

Khan (2017) used a survey-based method to investigate *YouTube* behaviors. Participants responded to questions about their uploading, liking, disliking, commenting, and sharing activity on *YouTube* on a Likert scale; questions were designed to address a variety of motives such as seeking information, social interaction, or relaxing and seeking entertainment. A regression of survey results showed that the social interaction had the largest coefficients for commenting; information seeking and giving information were also positively correlated with commenting on videos. Andersson (2021) considered impoliteness in comment threads for ten *YouTube* videos with negative words (e.g., *terrible* or *hysterical*) in their titles featuring climate activist Greta Thunberg. Using word2vec on the ~33,000 comments and ~500,000 words, she examined which words were closest to Greta in semantic space, finding that most of these words had negative evaluative content. The results were interpreted as an indication that impoliteness serves to consolidate similar views.

Overall, although several studies have considered the addressivity and interaction patterns of comments or used quantitative and predictive methods to explore aspects such as the relationship between metadata fields, few studies have compared the spoken discourse of videos and the discourse of the comments thereupon. In the next section, the methods used to evaluate the content and sentiment of videos as well as the comments on those videos are described.

3. DATA AND METHODS

The starting point for the analysis was transcripts of videos indexed in CoNASE (see Section 1), a 1.3-billion-word corpus of ASR transcripts of videos uploaded to the *YouTube* channels of municipalities and other local government entities in the US and Canada.[4] Much of the content of CoNASE consists of transcripts of public meetings of local councils in which local government and community issues are discussed, but other

[4] https://cc.oulu.fi/~scoats/CoNASE.html

content types, including interviews, sporting events, performances, and news reports are included.

In this study, only those videos which had comments in CoNASE were considered. Assessment of the content of the transcripts and the comments on the corresponding videos, summarization of the transcripts, and topic modeling of transcript content were undertaken in four principal steps, schematically illustrated in Figure 2.

First, after retrieval of all available comments, a sentiment score was calculated for each transcript and for each comment using the twitter-roberta-base-sentiment-latest transformer model (Camacho-Collados *et al.* 2022; Loureiro *et al.* 2022). Next, the transcripts were summarized into short texts, ranging in length from one to ten short paragraphs, using the distilbart-cnn-12-6-samsun model (Schmid 2023). This step was undertaken to create more coherent topics (see below). Topic modeling was then undertaken on the summarized content, using the BERTopic library (Grootendorst 2022). Finally, the sentiment scores, as values along a cline negativity-neutrality-positivity, were analyzed for the eight largest topics in terms of transcript and comment sentiment.

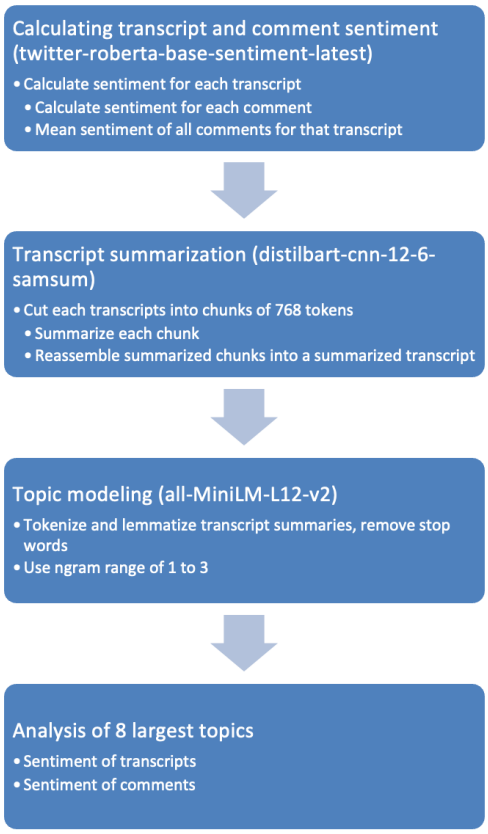

Figure 2: Schematic illustration of the processing and analysis steps. Transformer models are shown in parentheses

3.1. Transcript and comment retrieval and processing

The open-source *YouTube*-comment-downloader (Bouman 2022) was used to retrieve comments, via the innertube API, from videos whose transcripts are available in CoNASE. The vast majority of the videos in CoNASE have no user comments (and very few views), a fact which is unsurprising, considering the predictable nature of local government meetings and other municipal channel content. In total, of the 301,846 videos indexed in CoNASE, comments could be retrieved for 20,965. In addition, a small number of videos had been removed or made private (i.e., the comments were not available) in the time between the collection of the CoNASE data (2017–2021) and the time the comments were downloaded (mid-2022). The 190,097 downloaded comments ranged in length from 1 to 2,010 word tokens, with a mean value of slightly over 28 tokens.

3.2. Sentiment analysis

Sentiment analysis assigns negative, neutral, or positive sentiment to texts. Older, bag-of-words models, in which texts are assigned a value based on aggregate scores for individual lexical items, can perform poorly due to word order and contextual factors. A negative evaluation such as *he said it was great, wonderful, and fantastic, but it is really terrible* may be assigned a positive value based on the presences of three items with positive values and one item with a negative value. Similarly, language transformer models are typically better able to disambiguate the meanings of homonyms, determine the scope of negators, and correctly represent pronominal deixis due to their sensitivity to word-order and contextual considerations. A number of transformer-based sentiment analysis packages exist for text classification, but as *YouTube* comments tend to be rich in emoji, an analysis pipeline sensitive to emoji was selected, namely, the twitter-roberta-base-sentiment-latest transformer model (Camacho-Collados *et al.* 2022),[5] a fine-tuned sentiment model trained on 124m tweets (Loureiro *et al.* 2022), ultimately based on the RoBERTa pretraining approach (Liu *et al.* 2019).

While this model was appropriate for most of the comments in the data, which tend to be shorter in length, video transcripts are often much longer than the maximum input length for BERT models (often 512 tokens). An iterative procedure was therefore developed for texts longer than 512 tokens. They were converted to chunks of 512 tokens,

[5] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

then each chunk fed to the transformer model. The mean of the output vector values was then taken to be the sentiment for the entire text. The twitter-roberta-base-sentiment-latest model generates a vector of the likelihood of a given input being categorized as negative, neutral, or positive, outputting the argmax as a discrete value 0 (negative), 1 (neutral), or 2 (positive). This vector was converted directly to a continuous value in the range of 0 to 2 by calculating the dot product of the weighted values. For example, a model output of [0, .45, .55] indicates that according to the model, the text has zero percent probability of being negative, a 45 percent probability of being neutral, and a 55 percent probability of being positive. This corresponds to a score of 1.55, or mostly positive. The procedure was used to assign sentiment scores to all transcripts and comments in the study. An example is the video with the YouTube ID a1WSkvlw7zQ, entitled 'Meridian's new "Storey Bark Park"', uploaded by the YouTube channel of the City of Meridian, Idaho. The short video (1m 35s in length) consists of footage from the opening of a new community dog park, with a voice-over providing information on the event and about the park, and a speaker in the video making remarks at the opening ceremony. The calculated sentiment value for the transcript, which records a celebratory event, is 1.98, based on the overwhelmingly positive evaluative terms in the transcript (e.g., *huge success*, *enjoying the park*, *celebrate*). The single comment for the video is also positive and reads *Hands down the best dog park in the Treasure valley*, which was also assigned a value of 1.98.

3.3. Transcript summarization

Initial experimentation with topic modeling using the raw ASR transcripts produced inconsistent results. The BERTopic pipeline, which converts textual content to vector representations, is designed to process text with standard sentencization (i.e., periods or other sentence-ending punctuation) and is optimized for short texts such as sentences or paragraphs. The CoNASE transcripts with viewer comments which are analyzed in this study, however, have no sentence-ending punctuation, and vary greatly in length, from 100 to almost 80,000 tokens. To improve the quality of topics, a summarization step was undertaken for the video transcripts, using a recent transformer pipeline trained on transcripts of conversational speech (distilbart-cnn-12-6-samsum).[6] The model, based on

[6] https://huggingface.co/philschmid/distilbart-cnn-12-6-samsum

the BART architecture (Lewis *et al.* 2019), captures the essential discourse content of longer text passages and recapitulates it as short paragraphs.

First, transcripts were tokenized using *spaCy* (Honnibal *et al.* 2020) and split into 768-token chunks for summarization. The output for each transcript, consisting of 40-token summaries of the 768-token chunks, was then aggregated to generate the full summary for each transcript. The procedure reduced the variability in the length of the transcripts and introduced standard punctuation conventions. The resulting short texts, which retained the essential content of the longer transcripts, produced consistent topics, which upon manual inspection were found to correspond to most of the video content in the underlying video. For example, the video FUXTWgIqSfQ, entitled '"Sounds of Christmas" Christmas Band Concert', is a 47-minute recording of a school band performance. The transcript of the video, which is 1,649 tokens long, mainly consists of comments made by the band conductor. It comprises words of welcome and introduction to the audience, expressions of thanks to colleagues, parents, pupils, and band musicians, and introductions to each piece being performed. The summarized content of the video, which is 120 tokens long, foregoes expressions of welcome and thanks, beginning *The Bruton middle school intermediate band is playing the Nutcracker at the Bruton Christmas concert tonight.*

3.4. Topic modeling

Topic modeling (Blei *et al.* 2003) is an approach for the automatic identification of co-occurring word patterns, or topics, in sets of texts, which themselves can be defined in terms of the extent to which they participate in each topic. The technique, which can be considered a dimensionality reduction procedure, can be useful for the classification and interpretation of large sets of documents by distilling them into semantically interpretable topics. The default topic modeling approach utilizes relative word frequencies or term frequency-inverse document frequency values as input parameters for the algorithm. Traditionally, topic modeling is undertaken using 'bag-of-words' approaches based on word frequencies. While these can generate good results, they fail to account for sentence context. Transformer models such as BERT (Devlin *et al.* 2019), in which individual lexical items as well as immediate collocational contexts are represented by embeddings, or distributed vectors of numerical values, have been shown to be useful for a wide range of language processing tasks, including more robust topic modeling. This study utilized

BERTopic (Grootendorst 2022) for topic modeling. CoNASE transcripts were first summarized, as described above. Then, topic modeling was undertaken with all-MiniLM-L12-v2,[7] a model derived from miniLM (Wang *et al.* 2020), trained on 1.7 billion words of web texts from various genres and designed to map sentences and paragraphs to a multidimensional vector space for tasks like clustering or semantic search.

4. Results and analysis

4.1. Topics

The main input argument to the BERTopic algorithm is an array of textual content, in this case, a list of the 20,965 transcript summaries generated according to the procedure described above. In addition, the user can specify the underlying transformer model, the text tokenization procedure, the dimensionality reduction method, the words to be ignored (stopwords), and many other settings and parameters. For this analysis, tokenization was undertaken with the default CountVectorizer from scikit-learn (Pedregosa *et al.* 2011) and the English stopwords from NLTK (Bird *et al.* 2009). The output of the algorithm is the model, which can be inspected and visualized in many ways. One way to interpret the resulting topic model is to inspect the words which are most strongly associated with the topics in the model.

The eight largest topics, shown in Figure 3, represent the kinds of discourse that is typical for CoNASE transcripts, a large proportion of which are records of public meetings. The largest topic, Topic 0, is related to fire and rescue, services that are typically organized and funded by municipal governments in the United States. The provision of these crucial services often accounts for a considerable proportion of local government budgets, and discussion of, for example, hiring firefighters or the purchase of new equipment such as vehicles is a common discourse element in local government meetings. The words with the highest representation values in the topic include *firefighter*, *rescue*, *station*, and *department*, the latter two of which are likely collocates of *fire*, and *metro*, a term often appearing in the official names of municipal fire departments.

[7] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

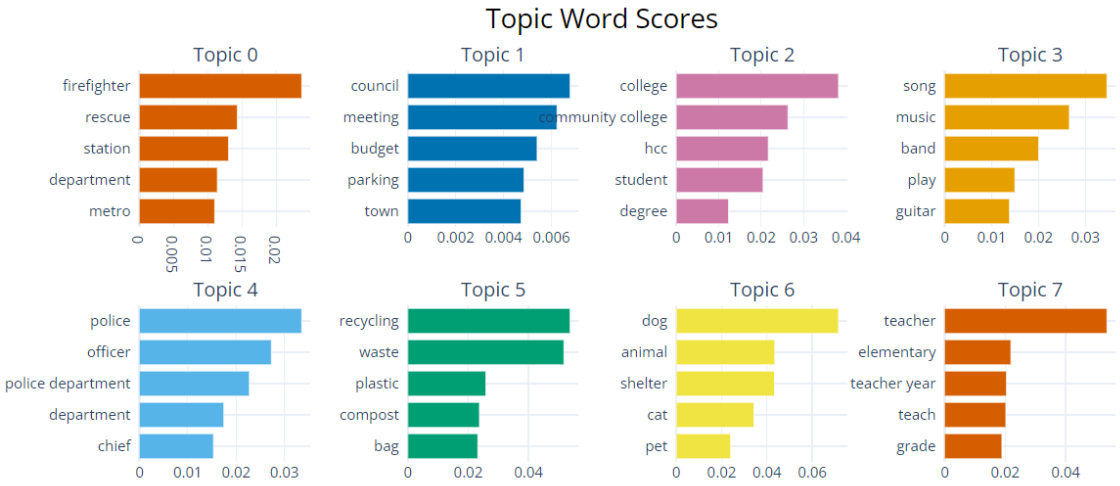

Figure 3: Words most strongly associated with the eight largest topics

Topic 1 includes words strongly associated with the content of municipal meetings: *council*, *meeting*, and *town* denote the activity of the municipal body itself, while *budget* and *parking* represent issues that are typical concerns of municipal governments. The items with the highest values in Topic 2 are from discourse pertaining to tertiary education: *college* and *community college* are places where the *student* can receive a *degree*. *Hcc*, in this context, is an initialism used to refer to several community colleges referenced in the discourse of CoNASE transcripts, including Houston Community College, Texas. In the United States, community colleges, which typically offer 2-year degrees, are often subsidized by municipalities. The videos from which the transcripts in this topic were taken include promotional content and interviews with community college presidents and staff members.

Topic 3 pertains to music. In the CoNASE corpus, it corresponds mostly to video transcripts of news announcements of upcoming musical performances, and occasionally of the performances themselves, for example those held during holiday or commemorative events, as well as performances organized by schools, universities, and other local organizations. The words most strongly linked to this topic denote music, instruments, and those who create music. Topic 4 represents another vital service of local governments. As is the case with fire and rescue, in the United States, most municipalities maintain a local police force and use local tax revenues for hiring and staffing the force as well as for procuring equipment such as uniforms and vehicles. The words in this topic denote police officers and the head of the police force, the *chief*.

Topic 5 deals with waste management, another service organized mostly by local governments. In addition to the words *trash* and *waste*, the words most strongly associated with this topic include *recycling*, *plastic*, and *compost*, indicating a concern for the environmental consequences of municipal waste and a desire to implement greener waste management policies. Topic 6 pertains to animals, as indicated by the words *animal*, *dog*, *cat*, and *pet*. The discourse for this topic relates to another service typically provided by municipalities in the United States and funded by local taxes, namely, animal control services, or the provision of facilities (in the form of a *shelter*) for stray and abandoned pets. Videos in the corpus with this topic include many in which animals at a shelter are introduced and offered for adoption. Topic 7 includes words used to discuss primary education, such as *teacher*, *teach*, and *grade*; *teacher year*, a word used in budgeting to describe the working hours of schoolteachers, and *elementary*, likely as a collocate of *school*. Primary education, in the US, is organized by municipalities and is therefore a frequent subject of discussion in municipal government meetings. In addition, the transcripts in this topic include content produced by school districts and schools themselves.

Overall, seven of the eight largest topics represent discourse that clearly pertains to local government decision-making: *firefighting*, *meetings*, *community colleges*, *police*, *waste disposal*, *animal control*, and *primary education*. These topics correspond to services that are provided at the local level by most municipalities in the US and Canada and whose concrete forms and budget allocations are the subject of much discussion by government representatives. The topic modeling procedure therefore accurately captures the fact that videos uploaded to municipal government channels are mostly about the immediate concerns of local governments, as captured in the discourse content of government meetings. In the next subsection, the sentiment expressed in those meetings, as well as in comments on the *YouTube* pages hosting those videos, are examined.

4.2. Sentiment

Both the transcripts and the comments in the data are more positive than negative, corresponding to the expected pattern for the sentiment of public discourse: communicators, in general, tend to accentuate positive sentiment and avoid expression of negative sentiment (Dodds *et al.* 2015). For this data, the mean sentiment value for transcripts was 1.20; for comments 1.29.

The distribution of sentiment values for transcripts and comments in Figure 4 shows peaks for comment sentiment near 0, 1, and 2. These peaks correspond to very short (mostly single-word or single-emoji) comments that are assigned a discrete value by the algorithm with a high probability score. Thus, a comment such as *great!!!* is assigned a value of 2 (positive), with 98 percent likelihood, whereas a comment such as *terrible!!!* or 🤮 would be assigned a value of 0 (negative) with high probability. As shown in Figure 4, single-token comments expressing positive sentiment are more common than neutral or negative single-token comments.

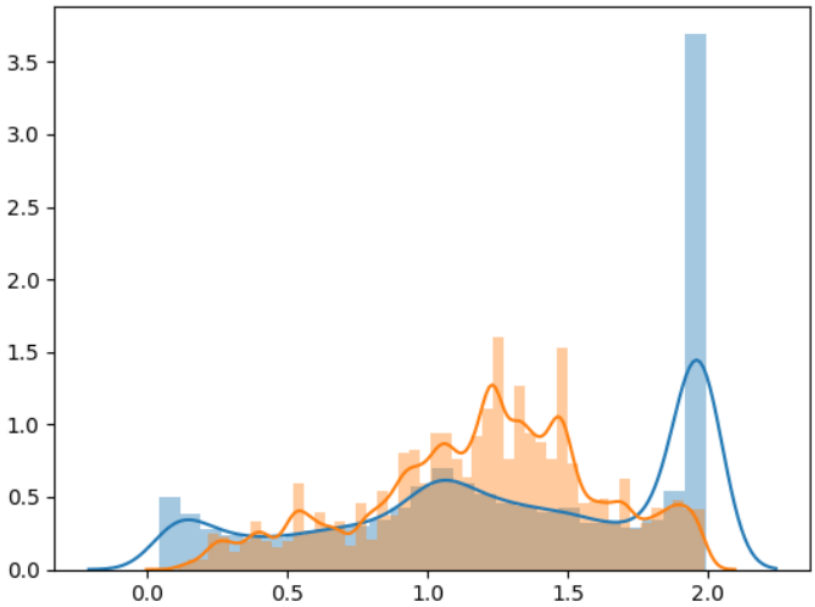

Figure 4: Distribution of sentiment values for transcripts (orange) and comments (blue)

The sentiment expressed in the video transcripts varies between the topics. Figure 5 depicts sentiment values for the eight largest topics in the transcript data. The median sentiment values for the topics, calculated on the basis of all the videos in the data assigned to that topic, range from 1.14, for the topic *meetings*, to 1.87, for the topic *school*. The sentiments expressed in the videos assigned to the topics *waste*, *firefighting*, and *police* have slightly lower median sentiment values of 1.23, 1.28, and 1.52. The topics *animal control*, *music*, and *community college* have higher median values: 1.54, 1.63, and 1.67.

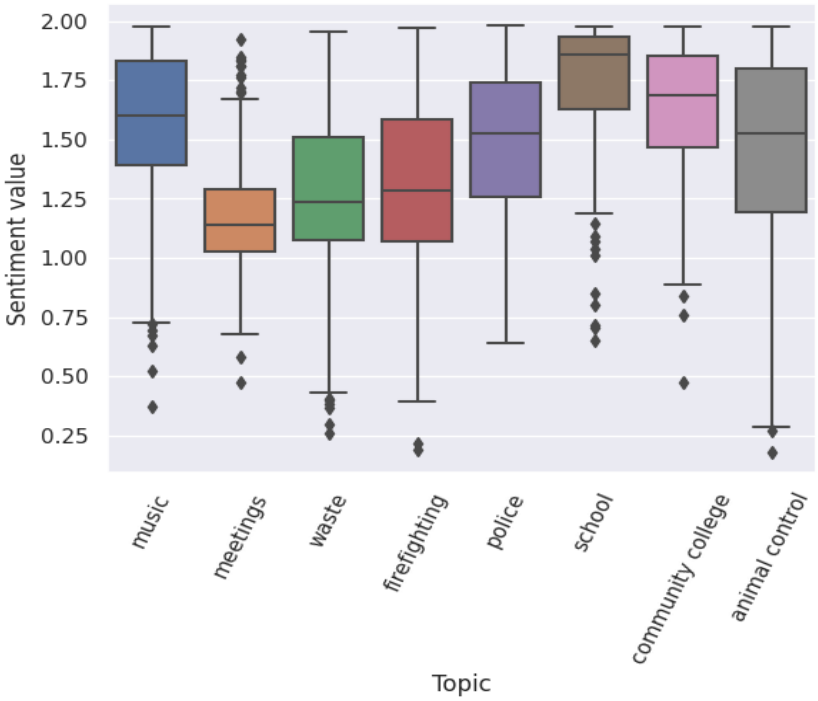

Figure 5: Transcript/summary sentiment for the eight largest topics

Comment sentiment, calculated as the mean for individual videos, tends to recapitulate the sentiment of the transcript summaries (Figure 6). For comments, median values per topic range from 0.93, for *meetings*, to 1.76, for *music*. The topics *police*, *waste*, and *firefighting* have median values of 1.08, 1.22, and 1.30, and the topics *school*, *animal control*, and *community college* median values of 1.50, 1.64, and 1.76.

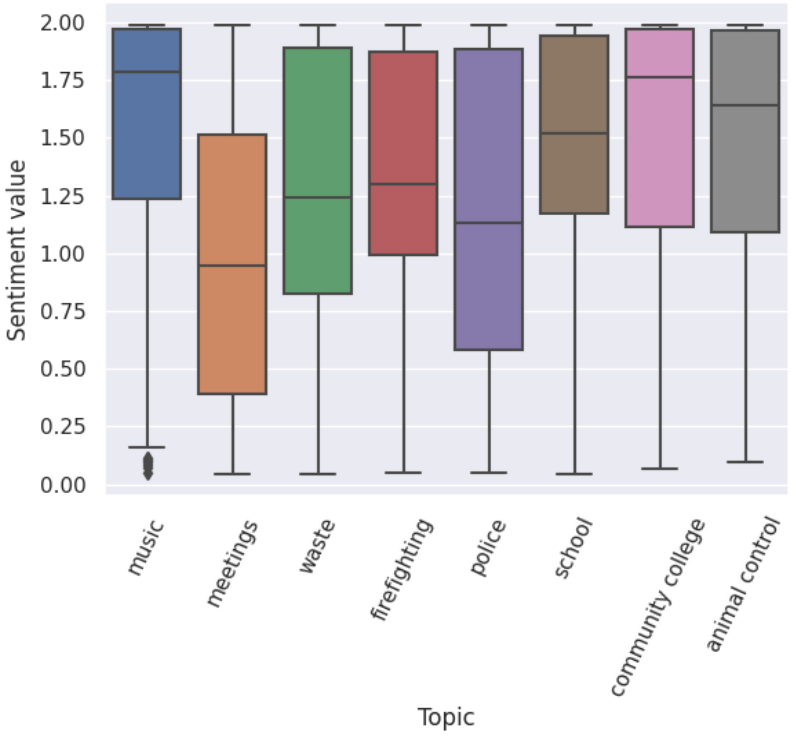

Figure 6: Comment sentiment for the eight largest topics

The difference between median sentiment values for the topics in terms of transcript content and aggregate comments may provide insight into the general contours of public perception of local government activities in the US and Canada.

Figure 7 shows the differences, per topic, between median comment sentiment value and median transcript sentiment values. Here, the topics that resonate positively with local communities become apparent: *music*, which as a topic exhibits positive sentiment on the basis of the transcript content, tends to attract comments that are even more positive. Likewise, transcripts with the topics *community college* and *animal control* attract comments that are more positive than the (already positive) sentiment contained in the transcript discourse.

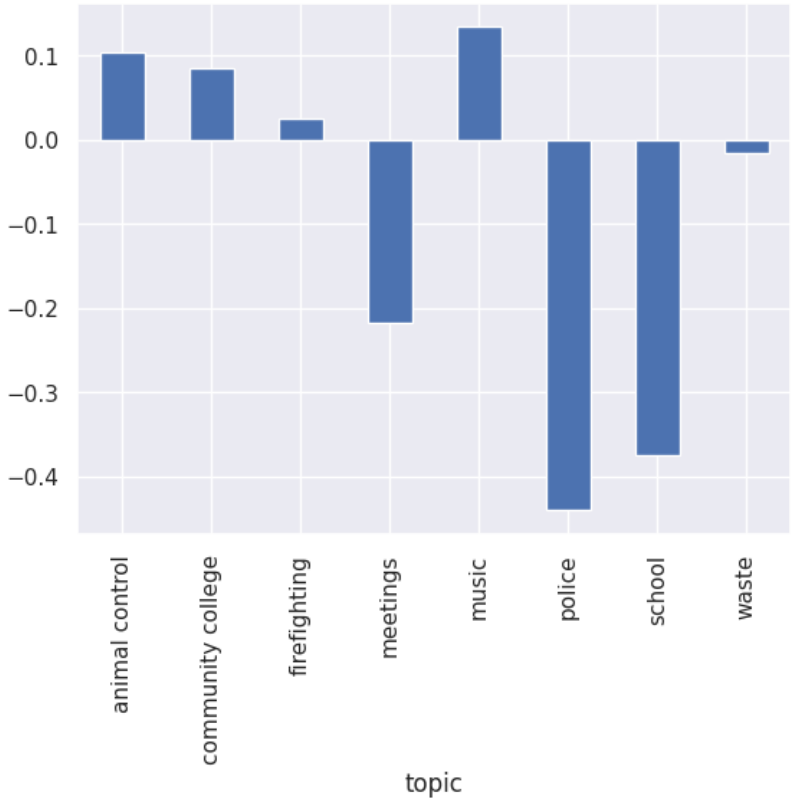

Figure 7: Difference in comment and transcript sentiment, per topic

The topics *firefighting* and *waste* attract comments that are approximately equivalent, in terms of median values, with the transcript content for those topics. Comments on *firefighting* are slightly more positive than the corresponding transcripts, while comments on videos with the topic *waste* are slightly more negative.

A different picture emerges for the topics *meetings*, *school*, and *police*. Here, the median sentiment of comments is significantly lower than the median sentiment for the videos. It is likely that the transcripts in the topic meetings are videos of local government

council meetings in which political decisions are discussed and debated, whereas transcripts in topics such as *music* or *animal control* include recordings of performances and informational videos about local organizations such as orchestras and animal adoption centers. The former type of video represents an interactive situation, both in the council chamber and in the comments section on the video's page, where critical and negative sentiments are more likely to receive expression. Disagreement is an important part of the political process, and council discussions are more likely to attract viewers who are critical of local government policies than are musical performances. Transcripts dealing with music and animals are more often informational, rather than discussion oriented. Videos showing musical performances or animals up for adoption are less likely to be criticized or discussed in a negative manner, not only because of their content, but because they are possibly less appropriate venues for the airing of disagreement.

Comment sentiment for the topic *school* is more negative than transcript sentiment, likely due to negative comments by pupils and parents. Comments on videos from this topic include remarks such as *School lunches suck* or *It's my opinion that [teacher name] gives way too much homework*, among others. While the topics *community college* and *school* both deal with education, school is sometimes perceived by pupils to be a burden imposed on them, against which *YouTube* comments may provide an opportunity for protest. Students at community colleges, on the other hand, choose to enroll in the college and usually must pay tuition fees, factors which may make them less likely to post negative comments.

The discourse pertaining to the topic *police* is also substantially more negative than the corresponding transcript material. Comments on the topic include general expressions of negative sentiment towards policing (*fuck the police*), as well as concern over overzealous and unprofessional policing practices, the awareness of which has increased in the last decade in the United States (comments for the topic include *unfortunately, lack of common sense and other far more disturbing behaviors with police officers seem to be commonplace* and *I beg you stop racial profiling it's evil and wrong racial profiling almost killed me*, among others). Commenting practices on videos pertaining to this topic appear to portray awareness of the fact that at community level, the practices and policies of many police forces in the US show room for improvement.

5. Discussion, outlook, and conclusion

It is perhaps unsurprising that popular sentiment is more positive towards topics such as *music*, *higher education*, or *pets*, compared to topics such as *waste management*, *meetings*, or *policing*. Higher education and music are universally acknowledged to be worthwhile and noble expressions of culture, and pets are objects of our love and affection. Furthermore, these topics represent areas where people can exercise agency: we choose to attend or view performances of music, to pursue higher education, and to own pets. Waste management, and policing, in contrast, are mostly perceived as external forces over which we have little influence, and which, in some cases, can be associated with unpleasant sensations, in the case of waste, or potentially dangerous situations, in the case of encounters with unprofessional police.

Although the automated methods of transformer-based sentiment analysis utilized in this study for gauging public attitudes may be new, they essentially recapitulate observations of previous generations towards music, perhaps most succinctly expressed by the English philosopher Herbert Spencer in 1854: “music must take rank as the highest of the fine arts—as the one which, more than any other, ministers to human welfare” (Spencer 2015 [1854]: 33).

The role of music as an uplifting and inspiring aspect of human existence, evident even in comments on *YouTube* channels of local governments, may have practical implications for the community outreach and engagement activities of municipalities. Local governments may be able to increase positive engagement with administrations by including content in their social media channels that reflects future-oriented aspects of communal life, such as education, music, and animals. In a broader perspective, the study represents an example of how transformer-based pipelines for text processing, including summarization, sentiment analysis, and topic modeling, can be used, in concert with ASR, to automatically gauge and assess aspects of communication and discourse.

Several caveats, however, should be noted, pertaining to the underlying data as well as the methods of analysis. The transcripts in the corpus contain ASR errors, with a mean Word Error Rate (WER) of approximately 15 percent (Coats 2024). Quality of ASR transcripts is influenced by both acoustic and dialect features, as highlighted in studies by Tatman (2017), Meyer *et al.* (2020), and Markl and Lai (2021). For this study, the sentiment analysis and summarization steps undertaken for the ASR transcripts ultimately rely on aggregate frequencies of word and n-gram types, as well as contexts. Although

inaccurate input data containing ASR errors may affect the precision of the results of these steps, it is unlikely to misrepresent overall trends in the data as long as the majority of automatically transcribed lexical items correspond to the correct types (see e.g., Agarwal *et al.* 2007).

The summarization step, in which long, unpunctuated ASR transcripts were converted to short paragraphs with standard punctuation, has not been validated for this kind of content (error-containing ASR transcripts). A validation of the accuracy of the summarization output for ASR transcripts would make the findings of the study more robust.

The topics generated by the BERTopic model are subject to a large number of variable input parameters, including the tokenization and lemmatization procedures for the text input, the underlying transformer (or other) architecture used to represent the input as numerical values, the algorithms for dimensionality reduction, as well as other parameters. Experimentations with various configurations of parameters showed that most input parameter settings resulted in the same largest topics. Nevertheless, the extent to which parameter variability can affect the model output, and hence the ensuing analysis, has not been assessed in this study.

The commenting behavior in the sample is not consistent. Some videos exhibit a very large number of comments, but most videos have just a few or one comment. A few comments are longer, in terms of number of tokens, but most comments are very short. This variability undoubtedly has an effect on the sentiment scores for the topics, and the significance of the calculated sentiment values has not been estimated. The method of comparing ASR transcript discourse with comment discourse, demonstrated in this study, may be better validated by selecting channels or videos with large numbers of comments. In addition, random sampling techniques for both videos and comments could help to demonstrate the relationship between transcript and comment content more robustly. In this respect —and considering the fact that municipal channel videos (such as those in CoNASE) typically have few comments, future studies, which do not necessarily need to consider engagement with local government— could target highly popular channels with extensive comments.

From a technical perspective, a few caveats should be remarked pertaining to the language models themselves. The twitter-roberta-base-sentiment-latest model, used to calculate comparable sentiment scores for transcripts and comments, was trained on

tweets, rather than on long, unpunctuated ASR transcripts. The accuracy of the model in predicting sentiment for longer texts remains unvalidated.

Despite these caveats, the study has demonstrated that large transformer models can be used in the context of computational social science for discovering the topical content of streamed or recorded meetings and for investigating the sentiment expressed therein, as well as for gauging the sentiment of comments on recordings of those meetings. While this finding has implications for media outreach for municipal governments or other kinds of organizations, the methods used in the study are not limited to analyzing organizational discourse. The potential utility of transformer models for research into communication and online interaction practices in general is great, and the comparison of speech content with commenting practices represents just the tip of the iceberg.

References

Agarwal, Sumeet, Shantanu Godbole, Diwakar Punjani and Shourya Roy. 2007. How much noise is too much: A study in automatic text classification. In Naren Ramakrishnan, Osmar R. Zaïane, Yong Shi, Christopher W. Clifton and Xindong Wu eds. In Naren Ramakrishnan, Osmar R. Zaïane, Yong Shi, Christopher W. Clifton and Xindong Wu eds. *Proceedings of the Seventh IEEE International Conference on Data Mining.* Los Alamitos; IEEE Computer Society. https://doi.org/10.1109/ICDM.2007.21

Andersson, Marta. 2021. The climate of climate change: Impoliteness as a hallmark of homophily in YouTube comment threads on Greta Thunberg's environmental activism. *Journal of Pragmatics* 178: 93–107.

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. Beijing: O'Reilly Media.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Bou-Franch, Patricia, Nuria Lorenzo-Dus and Pilar Garcés-Conejos Blitvich. 2012. Social interaction in YouTube text-based polylogues: A study of coherence. *Journal of Computer-mediated Communication* 17: 501–521.

Bouman, Egbert. 2022. *YouTube-Comment-Downloader*. https://github.com/egbertbouman/*YouTube*-comment-downloader

Camacho-Collados, Jose, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In Wanxiang Che and Ekaterina Shutova eds. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Abu Dhabi: Association for Computational Linguistics, 38–49.

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.

Coats, Steven. 2024. Noisy data: Using automatic speech recognition transcripts for linguistic research. In Steven Coats and Veronika Laippala eds. *Linguistics Across Disciplinary Borders: The March of Data*. London: Bloomsbury Academic, 17–39.

Coats, Steven and Veronika Laippala eds. 2024. *Linguistics across Disciplinary Borders: The March of Data.* London: Bloomsbury Academic.

Cotgrove, Louis A. 2022. *#GlockeAktiv: A Corpus Linguistic Study of German Youth Language on YouTube*. Nottingham: University of Nottingham dissertation.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran and Thamar Solorio eds. *Proceedings of 2019 Conference of the North American Association for Computational Linguistics: Human Language Technologies.* Minneapolis: Association for Computational Linguistics, 4171–4186.

Dodds, Peter Sheridan, Eric M. Clark, Suma Desu and Christopher M. Danforth. 2015. Human language reveals a universal positivity bias. *PNAS* 112/8: 2389–2394.

Dynel, Marta. 2014. Participation framework underlying YouTube interaction. *Journal of Pragmatics* 73: 37–52.

Gaventa, John and Gregory Barrett. 2012. Mapping the outcomes of citizen engagement. *World Development* 40: 2399–2410.

Goode, Luke, Alexis McCullough and Gelise O'Hare. 2011. Unruly publics and the fourth estate on YouTube. *Participations: Journal of Audience and Reception Studies* 8/2: 594–615.

Grieve, Jack, Dirk Hovy, David Jurgens, Tyler S. Kendall, Dong Nguyen, James N. Stanford and Meghan Sumner eds. 2023. *Computational Sociolinguistics*. Lausanne: Frontiers Media. https://doi.org/10.3389/978-2-8325-1760-4

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*: 2203.05794 [cs.CL]. https://doi.org/10.48550/arXiv.2203.05794

Häring, Mario, Wiebke Loosen and Walid Maalej. 2018. Who is addressed in this comment? Automatically classifying meta-comments in news comments. In Karrie Karahalios, Andrés Monroy-Hernández, Airi Lampinen and Geraldine Firzpatrick eds. *Proceedings of the ACM on Human-Computer Interaction*. New York: Association for Computing Machinery, 1–20.

Herring, Susan and Ashley R. Dainas. 2017. "Nice picture comment!" Graphicons in Facebook comment threads. In Tung X. Bui and Ralph Jr. Sprague eds. *Proceedings of the 50th Hawaii International Conference on System Sciences*. Hawai: University of Hawaii at Manoa, 2185–2194.

Herring, Susan and Seung Woo Chae. 2021. Prompt-rich CMC on YouTube: To what or to whom do comments respond? In Dan Suthers and Ravi Vatrapu eds. *Proceedings of the 54th Hawaii International Conference on System Sciences*. Hawai: University of Hawaii at Manoa, 2906–2915.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Khan, M. Laeeq. 2017. Social media engagement: What motivates user participation and consumption on YouTube? *Computers in Human Behavior* 66: 236–247.

Krohn, Rachel and Tim Weninger. 2019. Modeling online comment threads from their start. *arXiv*: 1910.08575v1 [cs.SI]. https://doi.org/10.48550/arXiv.1910.08575

Ksiazek, Thomas B. 2018. Commenting on the news. *Journalism Studies* 19/5: 650–673.

Ksiazek, Thomas B., Limor Peer and Kevin Lessard. 2016. User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media & Society* 18/3: 502–520.

Lehti, Lotta, Johanna Isosävi, Veronika Laippala and Matti Luotolahti. 2016. Linguistic analysis of online conflicts: A case study of flaming in the Smokahontas comment thread on YouTube. *Wider Screen* 19. http://widerscreen.fi/numerot/2016-1-2/linguistic-anaead-on-*YouTube*/

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv*: 1910.13461 [cs.CL]. https://doi.org/10.48550/arXiv.1910.13461

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*: 1907.11692 [cs.CL]. https://doi.org/10.48550/arXiv.1907.11692

Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. *arXiv*: 2202.03829v2 [cs.CL]. https://doi.org/10.48550/arXiv.2202.03829

Markl, Nina and Catherine Lai. 2021. Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation. In Su Lin Blodgett, Michael Madaio, Brendan O'Connor, Hanna Wallach and Qian Yang eds. *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing* . Association for Computational Linguistics, 34–40. https://aclanthology.org/2021.hcinlp-1.6

Meyer, Josh, Lindy Rauchenstein, Joshua D. Eisenberg and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 6462–6468.

Nycyk, Michael. 2012. *Tensions in Enforcing YouTube Community Guidelines: The Challenge of Regulating Users' Flaming Comments*. Perth, Australia: Curtin University of Technology dissertation.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12: 2825–2830.

Schmid, Hans-Jörg. 2020. *The Dynamics of the Linguistic System: Usage, Conventionalization, and Entrenchment*. Oxford: Oxford University Press.

Schmid, Phillip. 2023. *Distilbart-cnn-12-6-samsum*. https://huggingface.co/philschmid/distilbart-cnn-12-6-samsum

Schultes, Peter, Verena Dorner and Franz Lehner. 2013. Leave a comment! An in-depth analysis of user comments on YouTube. In Rainer Alt and Bogdan Franczyk eds. *Proceedings of the 11th International Conference on Wirtschaftsinformatik*. Leipzig: University of Leipzig, 659–673.

Siersdorfer, Stefan, Sergiu Chelaru, Jose San Pedro, Ismail Sengor Altingovde and Wolfgang Nejdl. 2014. Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web* 8/3: 1–39

Spencer, Herbert. 2015 [1854]. The origin and function of music. In John Shepherd and Kyle Devine eds. *The Routledge Reader on the Sociology of Music*. London: Routledge, 27–34.

Tatman, Rachel. 2017. Gender and dialect bias in YouTube's automatic captions. In Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube and Hanna Wallach eds. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia: Association for Computational Linguistics, 53–59.

Tausczik, Yla R. and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29/1: 24–54.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems* 33: 5776–5788.

Corresponding author
Steven Coats
University of Oulu
Faculty of Humanities
Pentti Kaiteran katu 1
Linnanmaa
P.O. Box 8000.
90014 Oulu
Finland
E-mail: steven.coats@oulu.fi

received: November 2023
accepted: February 2024

RiCL Research in Corpus Linguistics

Lost in a sea of highlight reels: The use of social media and mental health metaphors in online health blogs

Jennifer Foley
Autonomous University of Madrid / Spain

Abstract – This article explores the metaphorical conceptualisation of social media and its relationship with mental health and well-being in a specialised corpus of online health blogs, with the aim of discovering how people communicate their experience of social media use, and whether it has a positive or negative influence in their lives. For this purpose, a 20,000-word corpus of blog posts from online health communities, charities and personal blogs were collected and analysed. The main research questions are: a) How is social media conceptualised? b) Are metaphors used to conceptualise social media evaluative? c) How are mental health and well-being conceptualised? d) How are metaphors used to discuss the benefits and challenges of social media use for individuals who suffer from illness? Results show that the DRUGS, PLACE, PATH and FOOD source domains are used to conceptualise social media, and that metaphor is used to highlight both the positive impact of social media in providing social support and its negative impact on symptoms when used excessively.

Keywords – social media; metaphor; mental health; well-being; evaluation

1. INTRODUCTION

The influence of social media on mental health and well-being is currently a highly controversial issue, as some people claim it is detrimental to their mental health while others say it plays a vital role in their daily lives. Despite being widely studied in fields such as psychology and anthropology (Keles *et al*. 2020; Miller *et al.* 2021), few studies have investigated how this topic is communicated through linguistic choices in real world data, such as blogs or mediated communication (for exceptions, see Naslund *et al*. 2014). Furthermore, while studies on the evaluative potential of metaphor have frequently demonstrated that metaphor is an extremely useful tool in the communication of severe mental disorders (Coll-Florit and Climent 2022), mental illnesses (Semino 2008; Charteris-Black 2012), and diseases such as cancer (Semino *et al*. 2018), little research exists on the relationship between social media and mental health.

Research in Corpus Linguistics 13/1: 26–56 (2025). Published online 2024. ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.13.01.03

Against this background, this study aims to analyse how social media is conceptualised in a specialised corpus of blog posts from online health communities, mental health charities, and mental health bloggers. Furthermore, it also aims to discover whether the evaluative potential of metaphor is used when discussing the benefits and challenges of social media use, and whether people view social media as a positive or negative influence on their mental health and well-being. To achieve these aims, I will address the following four research questions:

RQ1: How is social media conceptualised in terms of the target domains a) social media platforms, b) social media content, and c) social media use?

RQ2: Are metaphors used to conceptualise social media target domains evaluative? If so, what is the predominant value?

RQ3: How are mental health and well-being conceptualised?

RQ4: How are metaphors used to discuss the benefits and challenges of social media use for individuals who suffer from illness?

The paper is organised as follows. Section 2 presents the background to the study while Section 3 discusses the data and methodology used. Section 4 presents the qualitative and quantitative results from the data by using descriptive statistics. Finally, Section 5 offers some conclusions and addresses the limitations of the study and areas for future research.

2. BACKGROUND

2.1. Conceptual metaphor theory and the evaluative function of metaphor

This study is grounded in conceptual metaphor theory (henceforth, CMT), which views metaphors as tools to communicate complex and abstract entities or ideas in terms of more concrete and tangible ones (Lakoff and Johnson 1980). During this process, real or perceived qualities and attributes of a source domain are mapped onto a target domain, "so that we can see, experience, think and communicate about one thing in terms of another" (Demjén and Semino 2017: 1). As well as facilitating communication by helping people to explain complex experiences, metaphors also carry out an evaluative function by highlighting certain aspects of target domains while, at the same time, backgrounding others (Semino 2021: 51).

In her corpus study, Deignan (2010: 363) identifies "four mechanisms that speakers use to evaluate through metaphor: creating entailments, exploiting scenarios, choosing significant source domains, and mapping connotational meaning." The evaluative potential of metaphor has recently been demonstrated in several analyses, such as in Porto's (2022) investigation on the use of WATER metaphors in the Spanish press to discuss Syrian migration, Hidalgo-Downing and Pérez-Sobrino's (2023) study on Brexit metaphors in British newspapers, and Fuoli *et al.*'s (2022) examination of metaphors in a corpus film reviews.

2.2. Social media and metaphor

To date, little research exists on the metaphorical representation of social media, which is surprising considering its abstract nature. Social media platforms carry out several functions which provide a rich site for the production of metaphor, such as communicating with others, uploading and sharing content, taking part in online events, and participating in online communities.

Among research carried out on real world data, i.e., data generated without researcher interference, a recent study by Foley and Hidalgo-Downing (2024) found that journalists in a 10,000-word sample of British newspaper opinion articles employed the PERSON, PLACE, DRUGS, OBJECT, WAR, COMPETITION and JOURNEY source domains to conceptualise social media platforms and their use. In addition, le Roux and Parry (2020: 189) suggest potential metaphors that may be used when discussing social media use and its effect on mental health and well-being, which are based on metaphors they frequently utilised in seminars and lectures, e.g., *Social media is a Townsquare*.

Regarding digital environments, Girón-García and Esbrí-Blasco (2019) demonstrate that cultural knowledge about supermarkets influences the conceptualisation of digital frames, such as *Amazon* 'departments', and regarding digital society, Katzenbach and Larsson (2017) provide a dossier of articles that examine the implications of using certain metaphors that 'pervade' discussions on digital transformation in politics, culture, and economics. Finally, and perhaps more relevant to this study, previous research on the use of metaphor in the conceptualisation of the internet has revealed that metaphors change and evolve along with technology, e.g. The 'information superhighway' metaphor of the early 1990s that conceptualised the sharing and receiving

of information online has now become obsolete, as the internet, and thus social media, has become a place that provides opportunities for people to build communities and gather online (Isomursu *et al*. 2007).

2.3. Mental health, illness and metaphor

Given the extensive amount of research carried out in the field of CMT in the past decade on the metaphorical representation of illness, both mental and physical, it is surprising that social media's effect on mental health has not yet been addressed. Research on the use of metaphor in the communication of mental illness (Semino 2008; Charteris-Black 2012; El Refaie 2014; Tay 2017; Coll-Florit *et al*. 2021; Forceville and Paling 2021) has demonstrated that metaphor is an invaluable tool when communicating the difficult and subjective experience of living with mental illness. Furthermore, recent studies on the use of metaphor for the emotional experiences of pregnancy loss (Littlemore and Turner 2019) and living with advanced-stage cancer (Semino *et al*. 2018) have provided healthcare practitioners with insights and best care practices to support people through this difficult time.

2.4. Social media use, mental health, well-being and blogging

Currently, there is a vast amount of research available on the relationship between social media use, mental health and well-being. However, as this relationship is a highly complex and nuanced one, researchers and psychologists have investigated this topic from various perspectives and approaches, leading to conflicting results. A systematic review of the literature on social media's influence on depression and anxiety in adolescents (Keles *et al*. 2020: 90) found that while it is "fair to say that there is an 'association' between social media use and mental health problems," concerns were raised regarding the cross-sectional nature of the studies, which were mostly quantitative rather than qualitative.

Similarly, a meta-analysis carried out on the conceptual and operational approaches to computer-mediated communication and mental health (Meier and Reinecke 2020: 32) found that although results "suggest an overall (very) small negative association between social media use and mental health," more rigorous approaches are needed, i.e., studies must assess the quality of social media use rather than the quantity. Finally, in a review

of the literature on social media use and well-being, Kross *et al*. (2021) arrive at a similar conclusion, calling for more experimental and longitudinal studies rather than cross-sectional ones, as well as a validation of the methodologies (self-report questionnaires) that are used to study social media's impact on well-being. Ultimately, Kross *et al*. (2021) highlight that social media's influence on well-being depends on both how and why people use it.

Despite the extensive research available on social media use and mental health and well-being, less research exists on the use of social media by those who suffer from serious mental illness, such as bi-polar disorder, schizophrenia and major depressive disorders. Naslund *et al*. (2020) provide a summary of the current research on the use of social media by individuals who suffer from mental illness, which also takes into account social media's impact on well-being. Of particular interest to this study is the identification of benefits and risks of social media use for individuals with mental illness. (For a detailed list of these benefits and challenges, see Naslund *et al*. 2020: 247).

The benefits include:

1) Facilitating social interaction: Individuals living with mental illness are at increased risk of social isolation due to impaired social functioning, as well as symptoms which may prevent or cause difficulties with face-to-face interactions. Social media facilitates social interaction as online communication does not require an immediate response or the use of non-verbal cues. Furthermore, people who suffer from mental illness may have less access to social support outside of family members or health care practitioners (Brusilovskiy *et al*. 2016), and social media helps them feel less socially isolated as they can interact with peers and access other social groups.

2) Access to peer support network: Online peer support provides opportunities for those with mental illness to share and receive strategies for coping with illness, as well as the ability to establish relationships and receive support from those who suffer from similar experiences.

The challenges involve:

1) Impact on symptoms: Studies have shown that prolonged or heavy use of social media can contribute to an increase in the symptoms of mental health and negatively affect well-being. For example, negative comparison on social

media was found to contribute to "risk of rumination and subsequent increases in depression symptoms" (Naslund *et al*. 2020: 249). Similarly, symptoms of anxiety are associated with prolonged social media use.

2) Facing hostile interactions: Individuals are at risk of being exposed to triggers or negative interactions on social media via comments or posts.

Similar to the ways in which social media use can have a positive impact on the symptoms of mental illness, by providing a site for social interaction and peer support, blogging has also been found to act as a therapeutic outlet for people who suffer from mental illness. Miller and Pole's (2010) analysis of the content and characteristics of 951 health blogs found that many health blogs tend to focus on topics that involve stigmatising illnesses or situations, such as "mental health, reproduction, HIV/AIDS, and disabilities" (Miller and Pole 2010: 1517). They suggest that by sharing aspects of their lives that they may usually keep hidden, health bloggers may experience a therapeutic outlet or cathartic release. Similarly, Hu's (2019) survey of 50 mental health bloggers also identified a positive therapeutic effect of blogging. By sharing their stories "to help their peers fight not only the disease, but the self-stigmatisation and fear" (Hu 2019: 118) bloggers can increase their sense of self-worth by perceiving themselves as helping others. The content and characteristics of the corpus compiled for this study appear to reinforce both Hu's (2019) and Miller and Pole's (2010) research, as many of the authors provide advice and coping strategies for dealing with social media's effect on mental health and well-being, indicating that bloggers experience a therapeutic impact both from sharing their stories and helping their peers.

As well as providing a summary of the current research on metaphor, social media, and mental health, in giving the background to this study I also hope to: 1) draw attention to the practical outcomes of research on metaphor and highly emotional and complex experiences, and 2) highlight that although there is an extensive amount of research available on social media's effect on mental health and well-being, there is a need for more studies that investigate the quality of social media use, rather than the quantity. Finally, more research is needed to discover how people who suffer from mental illness use social media, why they do so, and whether social media use improves or increases symptoms of mental illness.

3. DATA AND METHODOLOGY

3.1. Data

The data for this study consists of a 20,000-word specialised corpus compiled from posts featured on online health communities (e.g. tinybuddha.com), mental health charities (e.g., mentalhealth.org), and personal blog pages. Although 20,000 words may seem limited in terms of data sets, this specific corpus was compiled from a larger 100,000-word corpus for the purpose of testing the annotation method and identifying source domain categories. Regarding characteristics of the specialised corpus, it includes 19 texts from 19 authors, and the majority are written in American English. As some texts were published anonymously, it is difficult to estimate the average age and gender of the authors. Eligibility criteria included the primary focus of the post being social media's effect on a form of mental illness or aspect of well-being, and the search function on websites was used to identify posts containing the keyword *social media*.

Several ethical considerations were taken into account when designing this study, especially given the fact that mental illness can be a particularly sensitive topic. The primary decision was to compile the corpus from blog posts that were easily accessible in the public domain, such as those posted on the pages of bloggers or on the websites of mental health charities and health communities (as opposed to websites, platforms, and forums for which an account or membership is required). Regarding posts from mental health charities and health communities, I only included posts published on the story or blog section of websites, as people are required to submit these posts for editorial review. I acknowledge that it may not have been authors' intention for their posts to be used for research purposes, but have understood that by engaging in this process, authors are aware that their posts will receive more visibility and reach a wider audience.

Where contact information was provided, I emailed authors to explain the purpose of the study and to inform them that any identifying information would be eliminated or changed, and in cases where posts were published anonymously, I contacted the platform moderators. I received written consent to use five posts, including one request to reference the author of the post in the study, but I did not receive a response for the remaining 12 posts. Of the twelve who did not respond, six were sourced from websites that required authors to submit their posts for editorial review in order to be published, and six were sourced from the pages of bloggers who, upon further inspection, appear to have stopped updating their blog pages.

3.2. Method

A three-step annotation protocol to identify metaphorically used expressions and their evaluative potential was applied to the sample, and expressions were coded using *Microsoft Excel*. To reduce annotation bias and subjectivity, the sample was annotated and coded separately by another researcher and, in cases where disagreement occurred, the annotation was revised accordingly. In this section, I will describe the process behind each step and demonstrate how it was applied by using examples from the sample.

3.2.1. Identifying metaphorically used words

Steen *et al.*'s (2010) the *Metaphor Identification Protocol VU University Amsterdam* (MIPVU)[1] tool was used to identify linguistic metaphoric expressions that conceptualised the following target domains:

a) Social media platforms, including software such as algorithms.

b) Social media content, features, and forms of engagement, e.g., *likes*.

c) Social media use, e.g., frequency of use and ways of connecting or disconnecting.

d) Mental health and well-being, including mental illnesses such as depression or anxiety, emotions and feelings.

To identify linguistic metaphors, researchers must first read the text to gain a general understanding of the article, then reread it and identify potential metaphorically used expressions. Following this, words or expressions are coded as metaphorically used when the contextual meaning contrasts with the most basic meaning in dictionaries of reference, which were the *Macmillan*[2] and *Collins*[3] dictionaries in this case. According to Pragglejaz Group (2007: 3), the most basic meanings of words tend to be: 1) more concrete (what they evoke is easier to imagine, e.g., see, hear, feel, smell, and taste), 2) related to bodily action, and 3) more precise (as opposed to vague).

---

[1] http://www.vismet.org/metcor/documentation/MIPVU.html
[2] I am unable to provide the URL for entries for the Macmillan online dictionary, as the site was closed after the data had been collected and annotated. However, links to expressions from *Collins* online dictionary will be provided, as both were used to verify metaphoric expressions.
[3] https://www.collinsdictionary.com/

For example, in (1), the word *place* was identified as a linguistic metaphor by comparing the contextual meaning with the most basic dictionary entry for *place* in the *Collins* dictionary, which states that "a place is any point, building, area, town or country."[4] As social media is not a physical place, the contextual meaning does contrast with the most basic entry of the noun *place*, and the expression was coded as a metaphoric expression.

(1) Social media is an amazing place to connect with the world around us.[5]

3.2.2. Identifying source and target domains

Once the metaphoric expression is coded, the specific source domain must be identified. In many cases, the metaphoric expression itself provided the specific source domain. For example, in (2), social media content (memes) was conceptualised as SPICY FOOD:

(2) … spending hours liking the day's spiciest memes.

To identify the specific target domains, it is necessary to revise the metaphoric expression in its context of use. For example, in the case of the *spiciest memes*, memes were categorised under the general target domain of SOCIAL MEDIA CONTENT. Upon reviewing the metaphoric expression in its context, it was clear that not all content is conceptualised as SPICY FOOD, but only content that users spend *hours liking*, which implies that they find *spicy* memes interesting. As the third entry for *spicy* in the *Collins* online dictionary is "informal – suggestive of scandal or sensation,"[6] *spicy memes* were interpreted as those that are sensational or interesting. As a result, the conceptual metaphor INTERESTING SOCIAL MEDIA CONTENT IS SPICY FOOD was identified.

Finally, as this study is target domain-based, the general target domains of social media content, platforms, and use have already been identified, as well as the target domain of mental health and well-being. To identify 'overarching' or general source domains, I relied on previous research regarding metaphor and the internet, social media, mental health, and health (Isomursu *et al*. 2007; Semino *et al.* 2018; le Roux and Parry 2020 and Coll-Florit *et al*. 2021, respectively). I also relied on the *Master Metaphor List* (Lakoff *et al*. 1991) and the MetaNet wiki, which are catalogues of research on metaphor

[4] https://www.collinsdictionary.com/dictionary/english/place
[5] Nicholls, Kat. 2018. How to Take Care of Yourself Online (happiful.com)
[6] https://www.collinsdictionary.com/dictionary/english/spicy

studies that include source-target domain mappings and relevant examples. Once specific source and target domains were identified, conceptual metaphors were coded, e.g., SOCIAL MEDIA CONTENT IS FOOD, which is based on the conceptual metaphor IDEAS ARE FOOD (Lakoff *et al*. 1991: 84).

3.2.3 Identifying evaluative metaphorical expressions.

Evaluative metaphoric expressions were identified using criteria applied in the annotation procedure for evaluative stance and metaphor developed by Hidalgo-Downing and Pérez-Sobrino (2024) and Hidalgo-Downing *et al*. (2024). In this study, evaluative metaphoric expressions were marked as 'positive', 'negative', or 'both' depending on the connotations in the context of use. For a metaphoric expression to be coded as evaluative, the specific target domain had to be clearly identifiable in the text.

As in Martin and White (2005), evaluation in metaphoric expressions may be inscribed (explicit) or invoked (implicit). Inscribed evaluation occurs in (3), where the conventional metaphor SUFFERING FROM ILLNESS IS FIGHTING A WAR is used to conceptualise illness as an enemy in a battle, thus negatively evaluating this experience:

(3) It's a battle that I let few help me with.

In some cases, the inscribed polarity of evaluative metaphoric expressions was reversed when the context in which they were used elicited the opposite value. For example, in (4), the inscribed negative evaluation of *fight* is negated by surrounding context, which indicates the author's determination to overcome difficulties and maintain a positive attitude.

(4) I choose to be that person. To fight, to reflect and grow.

The conventional metaphor ILLNESS IS AN OPPONENT is often employed by people who suffer from illnesses such as depression, and the difficulty of living with illness is conveyed by referring to their *fight* or *battle* with the disease. However, in this case, the author casts themselves as an agent in their *fight* against depression, similar to cancer patients' use of this metaphor in Semino *et al*. (2018: 106–107), in order "to express a desire and effort to get better, and present patients themselves as active and determined."

Invoked evaluation occurs when, rather than condemning or praising the target, metaphor is used to imply a judgement. For example, in (5), there is an implied negative evaluation of spending too much time using social media to see what other people are doing, as this is conceptualised as watching people through the windows of their home.

(5) Social media is like a window into other people's lives. How you gonna live your life when you're out here peeping in windows?

The evaluation of target domains typically occurs when source domain connotations are mapped onto the target domain, such as the use of the DRUGS source domain in this sample. In (6), the metaphoric expression *fixes* evaluates the specific target domain SOCIAL MEDIA CONTENT negatively, as the most basic dictionary entry (in *Collins*) for the noun *fix* that contrasts with the contextual example is "an injection of an addictive drug such as heroin."[7]

(6) … maintain a balance of getting your social media fixes without the damaging effects.

However, when the DRUGS source domain is employed to conceptualise the target domain SOCIAL MEDIA USE, this does not always result in a negative evaluation. For example, the practice of stopping using social media for a period of time is conceptualised as *detoxing* in (7):

(7) I took one week away from social media to detox and clear my mind.

The most basic dictionary entry for detox in the *Collins* dictionary is:

> If someone who is addicted to drugs or alcohol detoxes, or if another person detoxes them, they undergo treatment which stops them from being addicted.[8]

Thus, in this case, the metaphoric expression *detox* was coded as positive, given that stopping being addicted to a substance carries a positive connotation.

Finally, there were instances where the evaluative connotation was ambivalent, and marked as 'both' positive and negative given the context of use. For example, metaphorically conceptualising social media as a sedative is positively evaluated in (8):

---

[7] https://www.collinsdictionary.com/dictionary/english/fix
[8] https://www.collinsdictionary.com/dictionary/english/fix

(8) … everyone is equally tired and frustrated yet sedated by the cool blue light of their phones.

Typically, sedatives are prescribed to treat anxiety, not tiredness or frustration, and the author later regrets the effects of ... *the aimless scroll, the blank looks, the lack of human connect* caused by social media in the same text.

4. RESULTS

The results are organised so that each subsection (4.1–4.4) deals with one of the four research questions the study aims to address. Sections 4.1 and 4.2 provide a descriptive analysis of social media source domains and evaluative metaphors (respectively), Section 4.3 offers a descriptive analysis of source domains used to conceptualise mental health, and Section 4.4. presents an analysis of how people use metaphors to discuss the benefits and challenges of social media use, as outlined by Naslund *et al*. (2020: 247).

4.1. Social media source domains

In this corpus, 209 metaphoric expressions are identified to conceptualise the target domains of social media content, social media platforms, and social media use. The results were quantified and presented below to discuss the percentage of metaphoric expressions that employ specific source domains within each target domain.

RQ1: How is social media conceptualised in terms of the target domains listed below?

a) Social media content.

b) Social media platforms.

c) Social media use.

4.1.1. Social media content

33 expressions (16%) are used to identify social media content. The most frequent source domains are DRUGS and SUBSTANCE, as shown in Figure 1. The ‘other’ category in this figure and in subsequent figures comprises instances of the same source domain appearing twice or less in the corpus.

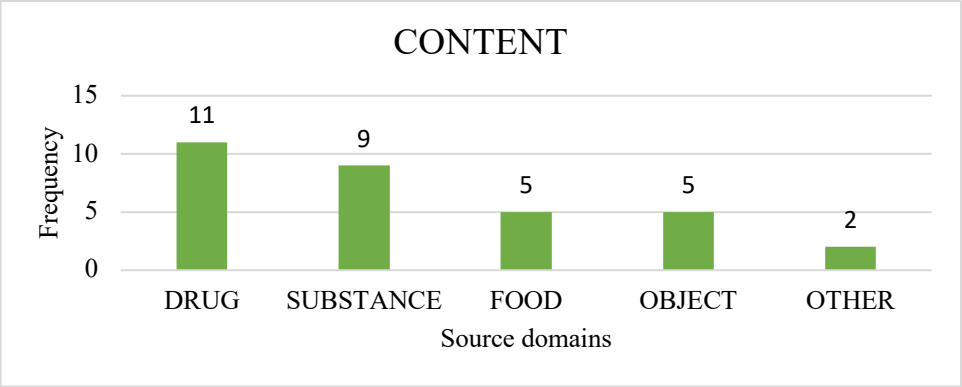

Figure 1: Social media content source domains

The DRUGS source domain is often used when people discuss their dependency on social media, either by using words that refer to *addiction*, as in (9), or by using words that are specific to drug use, as in (10):

(9) I had become addicted and consumed by Twitter.

(10) I'm a self-proclaimed social media junkie.

The SUBSTANCE source domain is used to conceptualise content that can *douse* (11) or *flood* (12) newsfeeds:

(11) No, I do not douse my social media in depressing posts.

(12) I've found when you are not surrounded by the constant flood of disheartening news stories …

The metaphor SOCIAL MEDIA CONTENT IS FOOD is employed to discuss social media content that people are *fed* (13), and to discuss content that negatively affects users (14):

(13) When our brains are fed news stories, social media feeds and email inboxes first thing in the morning …

(14) Think about what triggers you, what leads you to compare, whatever it is that leaves a bad taste in your mouth.[9]

4.1.2. Social media platforms

79 expressions (38%) are used to identify social media platforms. The most frequent source domains are PLACE, PERSON and OBJECT (see Figure 2).

[9] Nicholls, Kat. 2018. How to Take Care of Yourself Online (happiful.com)

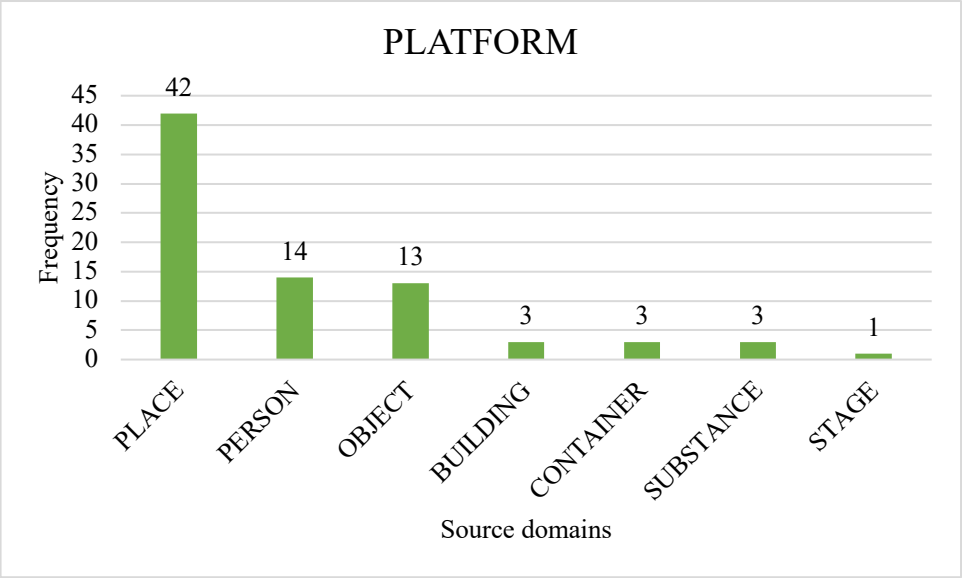

Figure 2: Social media platform source domains

The PLACE source domain is primarily activated by nouns that conceptualise social media as a physical place, such as *terrain* and *landscape* (15), or a *world* (16):

(15) … this technology is the new terrain on the landscape of communications

(16) It's easy to get lost in the beautiful place that is Instagram world.

Goatly (1997: 58) provides an in-depth list of how language constructs activities as places from which people can leave, enter, and move around in. In this sample, the PLACE source domain is also activated by verbs that foreground social media as a site for carrying out particular activities, such as *stalking*, as illustrated in (17):

(17) Before you know it, you have just spent 20 minutes stalking a total stranger.

The metaphor SOCIAL MEDIA IS A PERSON is employed when users conceptualise social media platforms as a person that can carry out certain actions, such as *taunting* people, as shown in (18):

(18) Social media can also taunt us by bombarding us with the adventures of people better left in our past.

Similarly, algorithms and platforms are also conceptualised as people that can be *trained* or *taught* to do something by changing settings to block certain content, as in (19):

(19) This trains the algorithm and teaches Instagram to show you more of the content you want to see.[10]

In addition, the OBJECT source domain is used when comparing accounts to *CVs* (20), or a *magazine* (21):

(20) Instagram is like your cool CV.

(21) … running a successful Instagram and blog is like running your own magazine.

4.1.3. Social media use

Finally, 96 (46%) metaphoric expressions are used to conceptualise social media use. The most frequent source domains are PATH, DRUG and FOOD (see Figure 3).

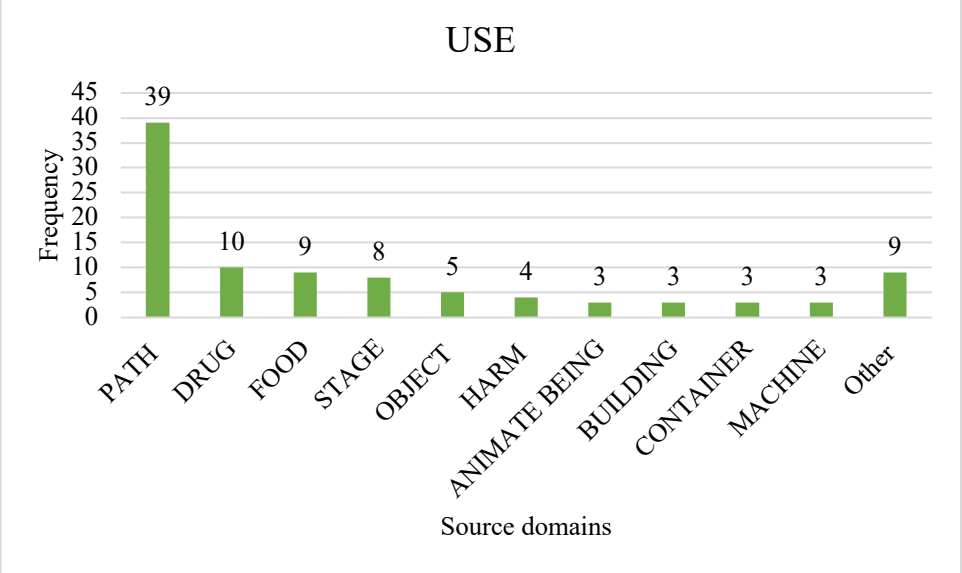

Figure 3: Social media use source domains

The PATH source domain is typically activated by verbs of motion, such as *hopping* (22) or *navigate* (23) between pages and websites:

(22) … hopping from one newsfeed to the next can be a good stress reliever.

(23) … it [social media] can be a fantastic and fun tool if I navigate and utilise it responsibly.

[10] Nicholls, Kat. 2018. How to Take Care of Yourself Online (happiful.com)

The PATH source domain is primarily used to discuss reducing or stopping the use of social media for certain periods of time. In these cases, although social media is conceptualised as a 'place', it is the movement to and from this place, i.e., connecting and disconnecting from apps, that is foregrounded. This is exemplified in (24) *vacation* and (25) *venturing back* below:

(24) I would also take vacations from social media by deleting social media apps off of my phone on the weekend.

(25) If your mood improves, then you can venture back in.

The DRUGS source domain is used when individuals conceptualise not using social media for a specific period as *detoxing* (26) or doing a *cleanse* (27):

(26) If not, it might be time to detox.

(27) … if that sounds like you, it's time for a social media cleanse.

Although a *cleanse* can refer to a variety of substances, such as unhealthy or harmful foods, it is interpreted as belonging to the DRUGS or FOOD source domain when other words that belong to that domain are previously activated within the text. For example, in the same article, the author discusses the addictive nature of social media and states that trying not to use social media is *like not being able to put down a cigarette or other addictive substance.*

Finally, the FOOD source domain is also used to discuss restricting social media use as going on a *diet* (28), and unrestricting this use for a certain period of time is conceptualised as a *cheat day* (29):

(28) I decided to go on a social media diet.

(29) A cheat day one Sunday afternoon (two hours of pure wasted social media time) left me feeling completely anxious.

The results in this section indicate that the target domain of social media use is the most productive site for metaphoric expressions, especially the PATH source domain. The prevalence of PATH metaphors is most likely motivated by the conceptualisation of social media as a place, which provides users with a means to discuss the action of connecting to or disconnecting from platforms, and the action of accessing different pages or accounts. It is interesting to note the trend in the evolution of metaphors regarding internet

use in the context of social media; while once people just 'surfed' the internet and exchanged information in the form of files using the 'information superhighway', the extensive amount of source domains employed to discuss the various aspects of social media and its use indicate that, nowadays, people rely on social media for much more than simply sharing information.

4.2. Evaluative social media metaphors

RQ2: Are metaphors used to conceptualise social media target domains evaluative? If so, what is the predominant value?

As can be seen in Figure 4, of the 209 metaphoric expressions identified, 135 (65%) are evaluative; 24 (18%) expressions are used to evaluate social media content, 39 (30%) to evaluate social media platforms, and 72 (53%) to evaluate ways of using social media.

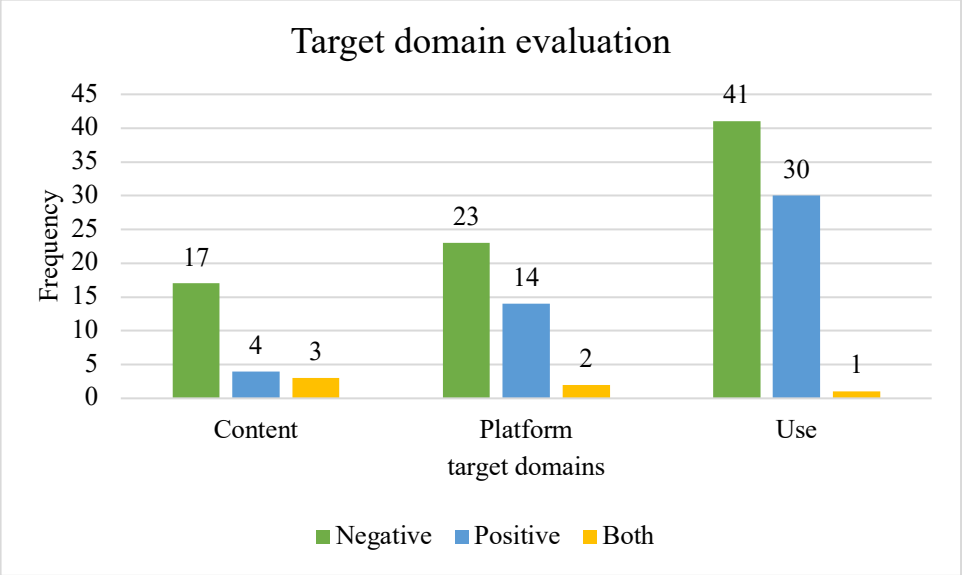

Figure 4: Evaluative metaphors for social media target domains

4.2.1. Social media content

Of the 24 metaphoric expressions that evaluate social media content, 17 (71%) are negative, 4 (17%) are positive, and 3 (12%) are coded as both positive and negative.

The DRUGS source domain primarily provides a negative evaluation of social media content when users discuss its addictive nature (30):

(30) Make no mistake about it, social media is addictive.

The SUBSTANCE source domain is used to conceptualise excessive negative content, such as *flood* (4; see section 3.2.3) or when content that stigmatizes mental health is seen as so abundant that it becomes difficult to 'wade through' (31):

(31) … we're left with a host of triggering and upsetting social media bumf to wade through.[11]

Regarding positive evaluation, the SUBSTANCE source domain is also used to refer to content that can be *sprinkled* over one's newsfeed, as shown in (32):

(32) Gone are the unnecessary reminders of particularly difficult moments, and at the top of my feed are sprinkles of humor and strength.

The instances where metaphoric expressions are marked as 'both' occur when the DRUGS source domain is used to discuss what is a potentially positive aspect of drug use, such as the 'sedative' example discussed in Section 3.2.3 (see example 8). Another example of this ambivalent evaluation in the sample occurred when one author described waking up to notifications in the morning as a *rush* (33) and something they looked forward to. However, the surrounding co-text also highlights how *addictive* this *rush* can be:

(33) It's like a little rush. Make no mistake about it, social media is addictive.

4.2.2. Social media platforms

Of the 39 evaluative metaphoric expressions that conceptualise social media platforms, 23 (59%) are negative, 14 (36%) are positive, and 2 (5%) are coded as both positive and negative.

The PERSON source domain negatively evaluates the way that algorithms decide what content appears on newsfeeds (34), and to conceptualise social media as a person with an *insatiable appetite* that *eats* your time, as illustrated in (35):

(34) Don't let the Facebook Wizard of Oz behind the curtain control how much support you get from people.

(35) If you're not careful, Facebook will eat your time. Its appetite is insatiable.

---

[11] Nicholls, Kat. 2018. How to Take Care of Yourself Online (happiful.com)

The OBJECT source domain is used to criticise the ways in which social media platforms negatively impact self-esteem, particularly when users compare themselves with others (36), or when people who suffer from illness negatively evaluate the way they *hide behind* social media accounts, because they worry that their symptoms will *embarrass* them when socialising, as illustrated in (37):

(36) It's a wacky funhouse mirror that distorts the image we see when we look into it.

(37) I hide behind my devices to avoid potential embarrassment, strengthening my anxiety in the process.

Regarding positive evaluation, the OBJECT source domain conceptualises social media as a *lifeline* that provides a way for people who suffer from illness to connect with others, as in (38):

(38) Do I leave the lifeline of social media instead?

Finally, the PLACE source domain is used when people discuss social media as an *amazing place* (39) that provides a way to connect with others, and to provide both a positive and negative evaluation of social media when one author highlights that there are positive aspects of social media, despite the fact that it can be overwhelming at times (40):

(39) Social media is an amazing place to connect with the world around us.

(40) Not all is bad in the world of social media, not when you can access support groups …

4.2.3. Social media use

Of the 97 metaphoric expressions used to evaluate social media use, 42 (57%) are negative, 29 (42%) are positive, and 1 (1%) is coded as both positive and negative.

The practice of passively using social media, i.e., viewing content and posts as opposed to engaging with others and uploading content, is negatively evaluated when people who do this are conceptualised as *ghosts*, as illustrated in (41):

(41) … become one of those Facebook ghosts that sees everything but is never evidenced to have been there.

The FOOD source domain negatively evaluates social media use when one author who, having been following a social media *diet*, had a *relapse* (42) after unrestricting social media use for a day:

(42) A relapse on social media left me feeling bad.

When discussing how comparing on social media can make people feel socially isolated, one author used the PATH source domain to discuss how people *fall into* (43) this habit, while another author used it to highlight how difficult it is to moderate the time they spend on social media (44):

(43) … we fall into the trap of comparing ourselves to others as we scroll through our feeds.

(44) Soon, you are sucked in, creepily scanning through pictures…it's a slippery slope.

Regarding positive evaluations of social media use, the DRUGS and FOOD source domains are employed when people share methods of reducing social media use, such as *detoxing* (26; see section 4.1.3) and *diets* (45):

(45) I stopped comparing myself to others. This happened by day two of the diet!

The PATH source domain is used when people conceptualise *leaving* social media for a period of time when they feel that the way they use it has become problematic, as in (24) and (25; see section 4.1.3)

Finally, controlling the types of content that you see on your newsfeed is positively evaluated by conceptualising social media accounts as homes. The TV presenter Marie Kondo, who is famous for teaching people how to organise and declutter their homes, is referenced in (46). The author uses the OBJECT source domain to conceptualise social media accounts and content that don't *bring joy* as clutter, and she encourages people to go *full Marie Kondo* and declutter their newsfeeds:

(46) … look at who you're following on social media and decide if they bring you joy…go full Marie Kondo on your social media accounts.[12]

[12] Nicholls, Kat. 2018. How to Take Care of Yourself Online (happiful.com)

In summary, metaphors that evaluate social media in this corpus are more often used to provide a negative evaluation, particularly in the case of social media content. This appears to be not only due to the addictive nature of social media content, but also due to the abundance of negative content that both stigmatises mental illness and triggers its symptoms.

4.3. Mental health and well-being metaphors

RQ3: How are mental health and well-being conceptualised?

Regarding mental health and well-being, the analysis identified 168 metaphors that conceptualise mental illnesses, such as depression and anxiety, the symptoms of illness and whether they are improving or worsening, and the emotional state of authors and their general well-being. The most frequently used source domains are UP/DOWN or DARK/LIGHT SCHEMA, WAR, JOURNEY, CONTAINER, MACHINE, ANIMATE BEING, OBJECT, SPLIT-SELF, and PLACE (see Figure 5).

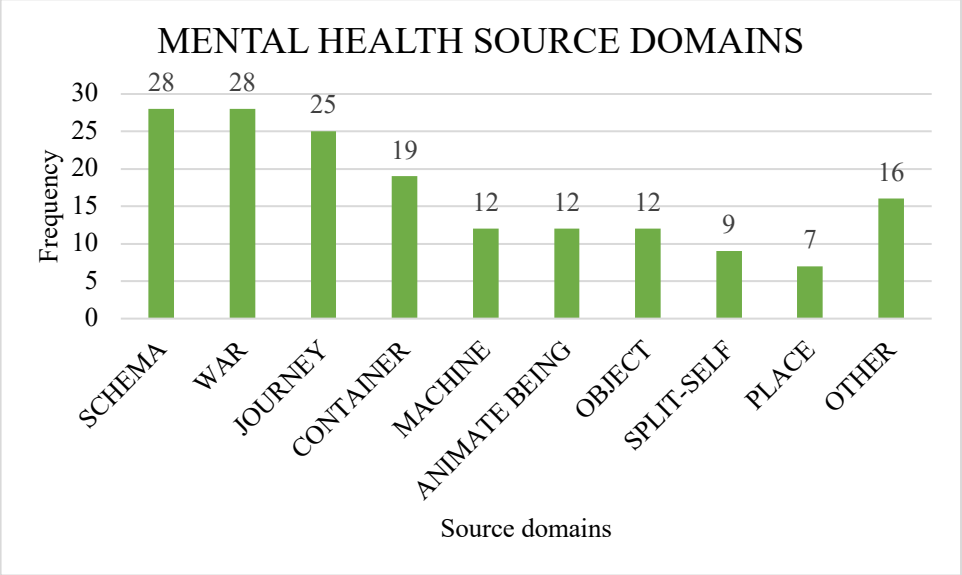

Figure 5: Mental health and well-being source domains

People often rely on image schema, such as GOOD IS UP to describe emotional states and well-being when they feel well or happy (47), or BAD IS DOWN when they are feeling unwell or sad, as shown in (48):

(47) When it's good, it's good—your self-esteem is high.

(48) But everyone has their low days.

The WAR source domain is highly conventional in illness discourse, particularly among people who suffer from depression. When people use this source domain, illness is conceptualised as an enemy that they must fight (49), and living with symptoms or going through treatment is conceptualised as an ongoing battle (50):

(49) When I realised this was affecting me, I choose[sic] to try to combat it.

(50) It's my struggle. It's a battle that I let few help me with.

As regards journey, the LIFE IS A JOURNEY metaphor is another highly conventional metaphor that people use to discuss living with illness, particularly when they conceptualise healing or getting better as forward motion along a path (51), or when they conceptualise symptoms that get worse as backwards motion along a path (52). While PATH metaphors conceptualise social media use as moving in and out of a place, JOURNEY metaphors highlight movement towards or away from a destination (goal).

(51) I look at what I can do to move towards the place I want to be.

(52) I feel like I'll be slipping back into nothingness and isolation.

The CONTAINER source domain is used to conceptualise a variety of experiences, such as the body as a container for emotions or energy that can be drained when using social media, as can be seen in (53). Furthermore, the CONTAINER source domain is also often used to discuss emotions or negative experiences as containers or bounded spaces that are difficult to *get out of* (54):

(53) Part of managing my health (as much as that's possible) is managing energy drains.

(54) I was in a funk, and it was hard to get out of it.

Regarding the MACHINE source domain, the PEOPLE ARE MACHINES metaphor is used to highlight how *unplugging* from social media can help people relax (55), or to discuss how people experience things differently because our minds are *wired differently* (56):

(55) Going unplugged for a few days can do wonders for your mental health.

(56) We are all wired differently; for some … social media … is soothing and provides solace.

As for the ANIMATE BEING source domain, people often conceptualise illnesses as animate beings in mental health and illness discourse, something which occurs in this sample when it is conceptualised as a *beast* that is difficult to escape from, as shown in (57):

(57) After wrestling with this relentless beast for more than 30 years, I have come to know its grasp …

Finally, as well as the conventional conceptualisation of illness as a burden or weight, the OBJECT source domain is also used to discuss people's reactions to, and engagement with, users' social media posts:

(58) It's also a tangled web of emotions.

In summary, the data discussed in this section provide further evidence that highly conventional metaphors such as LIFE IS A JOURNEY and the CONTAINER and WAR source domains appear frequently in discourse on mental health and well-being, thus contributing to the extensive existing research on the metaphorical conceptualisation of emotions, illnesses, and disorders.

4.4. Metaphors for the benefits and challenges of social media use

RQ4: How are metaphors used to discuss the benefits and challenges of social media use for individuals who suffer from illness?

In what follows, I provide a qualitative discussion on how people use metaphor to conceptualise the benefits and challenges of social media use. To present these results, I will rely on Naslund *et al*. (2020), who provide an in-depth discussion of the benefits and challenges of social media use for individuals with serious mental illness (see Section 2.3.).

4.4.1. Benefits

One of the benefits that social media use can provide for people who suffer from illness is its capacity to facilitate social interaction. This is significant, as individuals who suffer from illness are at risk of social isolation when symptoms prevent them from interacting with others. This benefit is demonstrated when people employ the PLACE source domain to conceptualise social media as a site they can *visit*, positively evaluating it as *the place*

where I can get out even when I am trapped inside. The social connection that platforms provide is crucial and is highlighted by one author who states that it *makes me feel like I exist when I feel myself fading away*.

Another author uses the SPLIT-SELF metaphor to highlight that social media interaction is a *critical part of what made me feel whole*, which is based on the PROPERTIES ARE POSSESSIONS conceptual metaphor. The author realised she had *lost* this part of herself when she stopped using social media for a period of time, and afterwards found that when it is used responsibly, social media *can become a place where mental health support and connection flourishes*. However, it appears that users should take care not to become dependent on social media for interaction, as this became a problem for one author when it resulted in her *surrender to one of the most harmful symptoms of social anxiety*, as relying on social media led to her avoiding face to face interaction.

Social media also provides access to peer support networks, which people often utilise to share tips and receive strategies for coping with illness. In this sample, the sharing of tips and strategies was identified when people employed the DRUGS and FOOD source domains to discuss the positive experiences of going on social media *diets* and *detoxes*. The practice of taking a break from social media use was also discussed using the PEOPLE ARE MACHINES metaphor, with disconnecting from social media conceptualised *as going unplugged for a few days in order to reset your mind*.

Another benefit of social media use is highlighted when people discuss receiving or providing support online; for example, the UP/DOWN schema is used by an author who states that talking about the *ups and downs* of sobriety online helped both themselves and others, while another mentioned that while feeling *low*, she supported others and *tried to lift people up*.

Finally, the LIFE IS A JOURNEY metaphor is used to discuss improvements in symptoms and how these were achieved, such as when one author stated that unfollowing people that negatively influence well-being was as a *great step* in her life.

4.4.2. Challenges

One of the challenges posed by social media use is that it can increase symptoms of mental illness, particularly when people engage in negative comparisons that *fuel*

insecurities. The SUBSTANCE source domain is used to conceptualise content when one blogger found that exposure to the achievements of others online was *drowning me instead of inspiring me*. Another author employed the SPLIT-SELF metaphor to discuss how comparing themselves to others online *began to tear me and my self-esteem apart*.

Naslund *et al*. (2020: 249) found that negative comparison contributes to the "risk of rumination and subsequent increases in depression." Rumination, which involves repeatedly thinking about or fixating on negative feelings and events, was specifically singled out as a consequence of comparing with others online, as shown in (59).

(59) My brain held a continuous whispering soundtrack called, "I'm not good enough."

Fortunately, some individuals are aware that comparing themselves to others online can negatively influence their well-being and they take steps to manage this. For example, one author pointed that *you have ultimate control over who you follow* and used the OBJECT source domain to conceptualise accounts that you can *get rid* of. Similarly, a blogger invited influencers to share their experiences of how comparing themselves online negatively affected them, in order to *pull back the curtain and let you know what's really up*. In this instance, the metaphor SOCIAL MEDIA IS A STAGE was employed to conceptualise followers as audience members, social media accounts or profiles as the stage, and the work that goes into creating posts and content as activity that occurs backstage.

Naslund *et al*. (2020) also state that prolonged use of social media can cause symptoms of mental health to increase and can negatively affect well-being. One individual stated that they stopped using social media because they felt like they were *being swallowed alive by the symptoms of mental illness*, while another author stated that they were *happy to be free from the burden* of social media but employed the FORCE schema to highlight how they felt themselves *being pulled to re-download the apps*.

While some did manage to reduce their social media use, other people commented that social media's addictive nature made it difficult to do so, employing idiomatic expressions that activate the PATH source domain to conceptualise unintentionally using social media for too long (60), or to express how difficult social media use is to moderate by comparing it to the moderating addictive substances (61):

(60) … between the allure of the endless scroll and the voyeuristic element, it's hard not to fall down a rabbit hole.

(61) Like trying to moderate alcohol, it's too much of a slippery slope.

The DRUGS source domain is used to discuss being unable to restrict *technology binges* that left one author with a *nagging sense of emptiness*, while another used the CONTAINER source domain to describe finding themselves in a *self-imposed prison of mindlessness*.

Another challenge of social media use is exposure to hostile interactions or triggers via comments and posts. For example, the WAR source domain is used to conceptualise social media *bombarding us with people better left in our past*. Similar harm can be caused by seeing content that triggers negative emotions and feels like *a bullet in the back*, leading to rumination when the post *left the confines of the screen and filled my room and my mind*.

For those who suffer from illness and had managed to become aware of situations that can trigger symptoms in their daily lives, the PEOPLE ARE MACHINES metaphor was used to highlight how social media posed a new challenge because they had to identify a *fresh set of switches* that could *cause my sleeping ogre to awaken*. While some people take steps to manage these triggers, acknowledging that they have to *tread carefully*, others decide to *leave* social media because of its potential to increase symptoms.

Finally, I will address a potential benefit/challenge of social media use that was not identified by Naslund *et al*. (2020,) but has been highlighted in the sample, namely escapism as a coping mechanism. When conceptualising social media as a place, one author praised being able to *visit worlds ways away from my own* when life became overwhelming. However, this was only a temporary solution to symptoms of illness, as the author stated eventually the *depressive thoughts return*. For some, symptoms became worse when social media was used as a coping mechanism, with one author stating that this means of escape can *quickly lead me down a rabbit hole of anxiety*. Finally, some individuals seem to be aware of the risk of using social media as a coping mechanism when experiencing symptoms of illness, stating that during these periods *a trip on social media is the worst thing*. In drawing attention to the potential negative aspects of using social media as coping mechanism, one author used the LIFE IS A JOURNEY metaphor to

highlight that spending too much time on social media will prevent one from moving towards their destination (goal), as the vast amount of content that is available online is easy to get *lost* in, as shown in (62):

(62) Are you spending precious life moments lost in a sea of highlight reels?

To conclude this section, it is important to highlight the potential advantages of applying corpus linguistics and CMT alongside studies in fields that investigate people's behaviour, thought processes, and emotions. Using Naslund *et al.'s* (2020) analysis as a guide, this study has identified metaphors to support claims that social media provides opportunities for interaction and peer support, and that social media poses a challenge for people when it triggers symptoms of mental illness and exposes them to hostile behaviour and content online. Furthermore, the analysis has also identified that some people use social media as a coping mechanism to 'escape' when feeling overwhelmed. While at times this can provide a form of instant relief, symptoms often return when people stop using social media, and more research is needed to understand how this form of 'escapism' can affect symptoms and patients over a prolonged period of time.

5. CONCLUSION

This research has contributed to the study of evaluative metaphors in health discourse on social media and its relationship with mental health and well-being, and it has also contributed to research on the figurative understanding of social media and mental health.

Regarding RQ1, the most frequent source domains for social media content are DRUGS, SUBSTANCE and FOOD; the most frequent source domains for social media platforms are PLACE, PERSON and OBJECT; and the most frequent source domains for social media use are PATH, DRUGS and FOOD.

Regarding RQ2, 65 per cent of metaphoric expressions that conceptualise social media are evaluative, and evaluative metaphors are primarily used to negatively evaluate the addictive nature of social media content, the way that algorithms decide what content users see, and the passive or excessive social media use.

As for RQ3, the most frequent source domains to conceptualise mental health and well-being are UP/DOWN schema, WAR, JOURNEY and CONTAINER. Finally, the answer to RQ4 is that the benefits of social media use for people who suffer from illness are

primarily highlighted by conceptualising social media as a place that provides an opportunity for social interaction when symptoms prevent face to face communication. In contrast, the challenges of social media use are highlighted when users discuss the tendency to compare their lives to others, and to discuss how excessive use of social media platforms can increase symptoms.

Finally, a limitation of this study that may be addressed in future research is that it does not employ inferential statistics. The aim of this pilot study was to identify frequently used source domains for social media and to discover if individuals use metaphor to discuss social media's effect on mental health and well-being. Future studies may carry out inferential statistics on a larger corpus to discover whether the higher percentage of negative evaluative expressions is statistically significant, and to compare it with other corpora (as in Fuoli *et al*. 2022).

Another limitation is that, while manual annotation can provide valuable insights into the evaluative function of metaphor, visualisation software can identify features of corpora such as collocations, clusters, keyword analysis and KWIC concordances, which would enrich this research and shed more light on social media's effect on mental health and well-being.

Should both of these limitations be addressed in future studies, the overall results could be compiled to produce a 'metaphor menu', similar to that produced from the results of Semino *et al.*'s (2018) research on metaphor and cancer. This menu could be used for personal or professional purposes, where patients are presented with a collection of metaphors that provide different perspectives on social media's impact on mental health and well-being, so that they can choose metaphors that resonate with them. For example, when discussing how interacting with a certain type of content can trigger symptoms, people could be encouraged to think of social media as food. In the same way that eating too much junk food in one sitting or too frequently can make us feel ill, frequently viewing or interacting with negative content can also cause us to feel ill.

In conclusion, this study has contributed to the under-researched area of the conceptualisation of social media by identifying which source domains people rely on to communicate their experience of social media use. In addition, the analysis has demonstrated that metaphor is a valuable tool for investigating the benefits and risks of social media use for mental health and well-being, as it provides a way of analysing this topic using real-world data (blogs and articles) instead of self-report questionnaires,

which can affect results due to bias. Finally, by approaching this topic from CMT, the highly complex relationship between social media use and mental health and well-being can be studied from a range of perspectives. Some of the perspectives are how people are influenced by specific content, how people evaluate the way that social media algorithms prioritise which content they see on their newsfeeds, and why people decide to leave social media platforms or return to them.

REFERENCES

Brusilovskiy, Eugene, Greg Townley, Gretchen Snethen and Mark S. Salzer. 2016. Social media use, community participation and psychological well-being among individuals with serious mental illnesses. *Computers in Human Behaviour* 65: 232–240.

Charteris-Black, Johnathan. 2012. Shattering the bell jar: Metaphor, gender, and depression. *Metaphor and Symbol* 27/3: 199–216.

Coll-Florit, Marta and Salvador Climent. 2022. Enemies or obstacles? Metaphors of war and journey in mental health discourse. *Metaphor and the Social World* 12/2: 181–203.

Coll-Florit, Marta, Antoni Oliver and Salvador Climent. 2021. Metaphors of mental illness: A corpus-based approach analysing first-person accounts of patients and mental health professionals. *Cultura, Lenguaje y Representación* 25: 85–104.

Deignan, Alice. 2010. The evaluative properties of metaphors. In Graham Low, Zazie Todd, Alice Deignan and Lynne Cameron eds. *Researching and Applying Metaphor in the Real World.* Amsterdam: John Benjamins, 357–374.

Démjen, Zsófia and Elena Semino. 2017. Introduction: Metaphor and language. In Elena Semino and Zsófia Demjén eds. *The Routledge Handbook of Metaphor and Language*. New York: Routledge, 1–10.

Forceville, Charles and Sissy Paling. 2021. The metaphorical representation of depression in short, wordless animation films. *Visual Communication* 20/1: 100–120.

El Refaie, Elisabeth. 2014. Looking on the dark and bright side: Creative metaphors of depression in two graphic memoirs. *Auto/Biography Studies* 29/1: 149–174.

Foley, Jennifer and Laura Hidalgo-Downing. 2024. Instagram is a ridiculous lie factory: Creative and evaluative metaphors of social media in a sample of newspaper opinion discourse. *Metaphor and the Social World* 14/1: 85–108.

Fuoli, Mateo, Jeannette Littlemore and Sarah Turner. 2022. Sunken ships and screaming banshees: Metaphor and evaluation in film reviews. *English Language and Linguistics* 26/1: 75–103.

Girón-García, Carolina and Montserrat Esbrí-Blasco. 2019. Analysing the digital world and its metaphoricity: Cybergenres and cybermetaphors in the 21st Century. *Cultura, Lenguaje y Representación* 22: 21–35.

Goatly, Andrew. 1997. *The Language of Metaphors*. London: Routledge.

Hidalgo-Downing, Laura and Paula Pérez-Sobrino. 2023. "Pushing Britain off the precipice": A CDA approach to negative evaluative stance in opinion articles on Brexit. In Juana I. Marín-Arrese, Laura Hidalgo-Downing and Juan Rafael Zamorano-Mansilla eds. *Stance, Inter/Subjectivity and Identity in Discourse*. Bern: Peter Lang, 201–226.

Hidalgo-Downing, Laura and Paula Pérez-Sobrino. 2024. Developing an annotation protocol for evaluative stance and metaphor in discourse. *Text and Talk* 44/2: 197–221.

Hidalgo-Downing, Laura, Paula Pérez-Sobrino, Laura Filardo-Llamas, Carmen Maíz-Arevalo, Begoña Núñez-Perucha, Alfonso Sánchez-Moya and Julia Williams Camus. 2024. A protocol for the annotation of evaluative stance and metaphor across four discourse genres. *Revista Española de Lingüística Aplicada* 37/2: 486–517.

Hu, Yifeng. 2019. Helping is healing: Examining relationships between social support, intended audiences, and perceived benefits of mental health blogging. *Journal of Communication in Healthcare* 12/2: 112–120.

Isomursu, Pekka, Rachel Hinman, Minna Isomursu and Mirjana Spasojevic. 2007. Metaphors for the mobile internet. *Knowledge, Technology & Policy* 20/4: 259–268.

Katzenbach, Christian and Stefan Larsson. 2017. *Imagining the Digital Society – Metaphors from the Past and Present.* https://www.hiig.de/en/imagining-the-digital-society-metaphors-from-the-past-and-present/

Keles, Betul, Niall McCrae and Annmarie Grealish. 2020. A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth* 25/1: 79–93.

Kross, Ethan, Philippe Verduyn, Gal Sheppes, Cory K. Costello, John Jonides and Oscar Ybarra. 2021. Social media and well-being: Pitfalls, progress, and next steps. *Trends in Cognitive Sciences* 25/1: 55–66.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live by*. Chicago: University of Chicago Press.

Lakoff, George, Jane Espenson and Alan Schwartz. 1991. *The Master Metaphor List*. Berkely: University of California.

Le Roux, Daniel B. and Douglas A. Parry. 2020. The Town square in your pocket: Exploring four metaphors of social media. In Marié Hattingh, Machdel Matthee, Hanlie Smuts, Ilias Pappas, Yogesh K. Dwivedi, Yogesh K Dwivedi and Matti Mäntymäki eds. *Responsible Design, Implementation and Use of Information and Communication Technology*. New York: Springer, 187–198.

Littlemore, Jeannette and Sarah Turner. 2019. What can metaphor tell us about experiences of pregnancy loss and how are these experiences reflected in midwife practice? *Frontiers in Communication* 4. https://doi.org/10.3389/fcomm.2019.00042

Martin, John and Peter R. R. White. 2005. *The Language of Evaluation*. London: Palgrave Macmillan.

Meier, Adrian and Leonard Reinecke. 2020. Computer-mediated communication, social media, and mental health: A conceptual and empirical meta-review. *Communication Research* 48/8: 1182–1209.

Miller, Daniel, Laila Abed Rabho, Patrick Awondo, Maya de Vries, Marília Duque, Pauline Garvey, Laura Haapio-Kirk, Charlotte Hawkins, Alfonso Otaegui, Shireen Walton and Xinyuan Wang. 2021. *The Global Smartphone: Beyond a Youth Technology. Ageing with Smartphones*. London: University College London Press.

Miller, Edward Alan and Antoinette Pole. 2010. Diagnosis blog: Checking up on health blogs in the blogosphere. *American Journal of Public Health* 100/8: 1514–1519.

Naslund, John. A., Ameya Bondre, John Torous and Kelly A. Aschbrenner. 2020. Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioural Science* 5: 245–257.

Naslund, John A., Stuart W. Grande, Kelly A. Aschbrenner and Glyn Elwyn. 2014. Naturally occurring peer support through social media: The experiences of individuals with severe mental illness using *YouTube*. *PLoS ONE* 9/10: e110171. https://doi.org/10.1371/journal.pone.0110171

Porto, Dolores M. 2022. Water metaphors and evaluation of Syrian migration: The flow of refugees in the Spanish press. *Metaphor and Symbol* 37/3: 252–267.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol* 22/1: 1–39.

Semino, Elena. 2008. *Metaphor in Discourse*. Cambridge: Cambridge University Press.

Semino, Elena. 2021. "Not soldiers but firefighters" – Metaphors and covid-19. *Health Communication* 36/1: 50–58.

Semino, Elena, Zsófia Demjén, Andrew Hardie, Sheila Payne and Paula Rayson. 2018. *Metaphor, Cancer and the End of Life: A Corpus-based Study*. London: Routledge.

Steen, Gerard. J., Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr and Tryntje Pasma. 2010. *A Method for Linguistic Metaphor Identification. From MIP to MIPVU.* Amsterdam: John Benjamins.

Tay, Dennis. 2017. Using metaphor in mental healthcare. In Elena Semino and Zsófia Demjén eds. *The Routledge Handbook of Metaphor and Language*. New York: Routledge, 371–385.

Corresponding author
Jennifer Foley
Autonomous University of Madrid
Faculty of Arts and Philosophy
English Studies Department
Campus de Cantoblanco
28049 Madrid
Spain
E-mail: jennifer.foley@estudiante.uam.es

received: November 2023
accepted: July 2024

RiCL Research in Corpus Linguistics

Emoji use by children and adults: An exploratory corpus study

Lieke Verheijen – Tamara Mauro
Radboud University / The Netherlands

Abstract – Emoji (e.g., 🤪✈️🧁) are increasingly used on social media by people of all ages, but little is known about the concept 'emoji literacy'. To investigate different age groups' emoji preferences, an exploratory corpus analysis was conducted using an innovative corpus-gathering method: children and adults were instructed to add emoji magnets to pre-constructed printed social media messages. The corpus (with 1,012 emoji) was coded for the number of emoji used per message, the type of emoji, their position and function in the message, and the sentiment they conveyed. Intuitions about emoji use turned out to be similar for children and adults, with greater use of facial emoji, emoji at the end of messages, emoji to express emotions, and emotional emoji to convey positive sentiment. Children's emoji preferences were studied in more detail. Results revealed that their age, gender, smartphone ownership, and social media use related to differences in the number, position, and function of the emoji used. The data showed that older children, girls, children with their own smartphone, and children using social media exhibited a more advanced and sophisticated use of emoji than younger children, boys, and children without smartphones or social media experience. This study constitutes an important first step in exploring children's emoji literacy and use.

Keywords – emoji; social media; computer-mediated communication; children; digital natives; emoji literacy

1. Introduction

Digital messages are becoming increasingly visual (Thurlow *et al.* 2020). Text-based computer-mediated communication (henceforth CMC) can nowadays be augmented with visual elements such as emoji, stickers, GIFs, memes, photos, and videos (Wang *et al.* 2019). Emoji in particular abound in personal CMC (Coosto 2020) and professional CMC (Dijkmans *et al.* 2020). These colourful small images cannot just present facial expressions (😛, 😢, 😉, 😆), similar to the more old-fashioned emoticons consisting of typographic characters (*:p*, *:'(*, *;)*, *XD*), but also all kinds of activities (⛹️), animals (🦚), objects (🧢), and symbols (☢️). The range of emoji available in the Unicode Standard (Unicode 2023) continues to expand, with currently over 3,700 emoji, including different genders, skin tones, and countless flags. In 2015, Oxford Dictionaries even pronounced the 'face with tears of joy' (😂) emoji as 'word' of the year, which testifies to the ubiquity

and salience of emoji in digital writing (Steinmetz 2015). Emoji are a striking aspect of contemporary online language, making them a highly interesting research topic. The body of academic literature on emoji is expanding, but research on children's (i.e., digital natives') use of emoji is generally lacking. The present paper will fill this research gap by reporting on a corpus analysis exploring how children use emoji. The aim of the study is thus to explore children's inclinations for using emoji (e.g., 😀🍕🙏🌸). The following two research questions are addressed:

RQ1: Do children use emoji differently than adults?

RQ2: Which demographic factors affect children's use of emoji?

2. Theoretical framework

2.1. Emoji as a multifunctional resource

Emoji are one of the visual elements that can make social media messages multimodal. They are small graphical images, also called 'graphicons' (Herring and Dainas 2017; Dainas and Herring 2021), which contain considerable visual detail. Previous studies have examined the utility of emoji as a digital resource, showing that they can fulfil numerous communicative functions in online writing by combining the roles of images, words, ideograms, nonverbal signals, and punctuation marks (Dürscheid and Siever 2017; Siebenhaar 2018; Tang and Hew 2018; Cohn *et al.* 2018, 2019; Beißwenger and Pappert 2019; Dürscheid and Meletis 2019). Prior work has revealed that emoji representing faces (😊😍😜), gestures (👍🙌✌), or people (👩‍👨‍👦) can compensate for the lack of non-verbal communication and paralinguistic cues in writing, can change the meaning or tone of a message, can express emotions, and can convey humour (Verheijen 2016; Evans 2017; Gawne and McCulloch 2019; Seargeant 2019). Other emoji (🎁🧁🐶) can simply visualise, 'decorate', or disambiguate text, thereby reducing chances of misinterpretation (Riordan 2017b). Emoji can make messages more playful or informal, indicating a sense of intimacy or social familiarity (Stark and Crawford 2015; Riordan 2017a). They can be used to structure messages, complementing or replacing punctuation marks (Dürscheid and Siever 2017; Pappert 2017; Busch 2021). In terms of speech acts, emoji can change the locution —the literal meaning of a message— and illocution —how the sender intends a message to be interpreted—, thereby affecting the perlocution —how a message affects

the recipient— (Austin 1962; Searle 1969). Drawing on Spina's (2018) work on emoticons, emoji can, in short, be designated as having semiotic, emotional, social, structural, and pragmatic functions.

Not everyone interprets emoji in the same way. Dainas and Herring (2021) point out that many emoji are semantically ambiguous. As previous research indicates, variability in emoji interpretations occurs both within and between digital platforms, in semantics (meaning) and sentiment (valence/tone/positivity), when presented in isolation or in the context of messages (Tigwell and Flatla 2016; Miller *et al.* 2016, 2017; Weissman 2019; Franco and Fugate 2020). Such a variation in emoji meanings also exists because besides a denotation (the literal/surface meaning), emoji can have multiple connotations (i.e., non-literal/figurative meanings), which may be metaphoric or euphemistic (e.g., 🐍🍆, Weissman 2019). Differences in emoji interpretations can be dependent on users' age, where younger people tend to be more familiar with novel connotations (e.g., 💀 to express dying from extreme laughter) and older people are more prone to 'incorrectly' interpret emoji (e.g., using 😂 in a sad context) (EditieNL 2016; Abril 2022).

Today's children are growing up with practically unlimited access to digital resources, whereas adults have only learned the ways of CMC at a later age. Younger generations, the 'digital natives', are more familiar with CMC —including emoji— than older generations, the 'digital immigrants' (Prensky 2001; Frey and Glaznieks 2018). Tailored to emoji, natives were born after emoji were invented in 1997. The present paper will study emoji usage by digital natives and digital immigrants from a multitude of approaches, including a) their semiotic use (by examining different types of emoji), b) their structural use (by examining different positions of emoji), c) their pragmatic use (by examining different functions of emoji), and d) their emotional use (by examining different sentiments of emoji).

2.2. Emoji literacy

In this digital day and age, the literacy landscape has been transformed up to the point where traditional literacy no longer suffices. Rather, a mastery of multiple literacies is required to succeed in society. Such new literacies include —but are not limited to— what have been named 'computer literacy', 'digital literacy', 'new media literacy', and 'visual

literacy' (see Verheijen 2018, for an extensive overview and discussion of new literacies). Emoji are a striking visual element of digital writing. Hence, emoji literacy (coined by Danesi 2016) can be considered a subtype of visual literacy. Wang *et al.* (2019) emphasise that digital visual literacy includes more than just emoji, since emoji are part of a wider inventory of graphicons which includes other visual means of expression such as emoticons, stickers, GIFs, and memes. Still, the present paper zooms in on emoji, because these have become so highly integrated into digital writing that they have been incorporated into the Unicode Standard, which encodes most of the world's writing systems (Unicode 2023).

According to Danesi (2016: 88), being emoji literate means that "semantic, syntactic, reinforcement, and conceptual aspects of the grammar interrelate with each other to produce the meaning behind (or underneath)" emoji. Freedman (2018) argues that emoji literacy has a cultural dimension, because they originated in Japan. Scheffler *et al.* (2022) observe that emoji literacy bears similarities to traditional (or 'linguistic') literacy. However, Freedman (2018) and Scheffler *et al.* (2022) focus on the comprehension of emoji, even though literacy crucially depends not just on reading but also writing skills, receptive and productive skills. As such, emoji literacy is determined by people's competence to read and write emoji, that is to say, to comprehend them and to use them. In this paper, we therefore define emoji literacy as the ability to understand and use emoji in appropriate ways in written CMC. Appropriate emoji use and understanding requires an awareness of different emoji meanings and a sensibility for differences in (online) registers.

Emoji literacy is key to effective digital writing. As Hurlburt (2018: 18, 15) rightly notes, "visual literacy, including the use of emoji, becomes an increasingly important skill" and emoji literacy needs to be acquired "to become a truly effective emoji communicator." Digital natives, who have grown up with digital communication tools and social media, can be expected to be more 'emoji literate' than digital immigrants, who have learnt to use such tools and media at a later age. Accordingly, digital natives have more positive attitudes towards emoji in general (Prada *et al.* 2018), are more familiar with (meanings of) emoji (Herring and Dainas 2020), and may be more proficient at attributing emotions to emoji. The current paper will explore if any differences in emoji use can be identified between digital natives and digital immigrants, and among digital natives (here, children) themselves.

2.3. Emoji and children

In recent years, emoji have come under increasing scrutiny of scientific research (see reviews by Bai *et al.* 2019; Tang and Hew 2019; Manganari 2021), but only little research has examined emoji perceptions or production by children. Research with a psychological approach has revealed that children can attribute emotions to facial emoji (Oleszkiewicz *et al.* 2017; Liu and Li 2021; da Quinta *et al.* 2023). Oleszkiewicz *et al.* (2017) found that children without social media or smartphone experience (between the ages of four and eight) can accurately interpret which emotions, especially happiness and sadness, are expressed by certain widely used facial emoji. This accuracy in emotion recognition from emoji was higher in girls and older children than in boys and younger children. Da Quinta *et al.* (2023) confirm that children (aged six to 12) can understand facial emoji. However, they add that such an understanding depends on the context of evaluation. Liu and Li (2021) sampled an even younger age group and showed that 30-month-old toddlers can already associate commonly used facial emoji with emotion words, thereby showing the first signs of emoji literacy.

In the field of education, previous research observed that emoji can also help children to understand emotions and other abstract concepts and to improve their self-expression (Fane 2017; Fane *et al.* 2018), that children can use emoji as storytelling devices (de la Rosa-Carrillo 2018), and that emoji can be used to measure children's attitudes to school subjects like mathematics (Massey 2022).

Most previous studies that have focused on children and emoji were in the domain of marketing and consumer research. Emoji on food packaging have been shown to affect children's dietary choices (Siegel *et al.* 2015; Luangrath *et al.* 2017). A substantial body of research has studied how emoji can be effectively utilised to measure children's emotional responses to food and other products (Gallo *et al.* 2017; Swaney-Stueve *et al.* 2018; Schouteten *et al.* 2018, 2019; Lima *et al.* 2019; Deubler *et al.* 2020; Sick *et al.* 2020a, 2020b; da Cruz *et al.* 2021; da Quinta *et al.* 2023).

Reviewing the relevant research that has been conducted on emoji thus far, it becomes apparent that children are a hitherto underexplored demographic in emoji research from a linguistic perspective. To our knowledge, this paper is the first pragmalinguistic study into children's emoji use, rather than into their perceptions or interpretations of emoji. The purpose of this study is twofold: a) to investigate if children (digital natives) use emoji differently than adults (digital immigrants) and b) to examine

which demographic factors affect children's emoji use. These questions will be addressed by analysing a corpus collected under semi-experimental conditions. The analysis will provide additional knowledge on emoji use by children as compared to adults and will thereby also contribute to existing theory on emoji literacy.

3. METHODOLOGY

3.1. Materials: Data collection

The research questions were addressed with a corpus collected at the *Kletskoppen Kindertaalfestival* in 2020, a language festival in Nijmegen (the Netherlands) aimed at children. The data were collected at this festival by means of "The Great Emoji Experiment" (*Het Grote Emoji Experiment*). 30 children (mean age = 8.5; age range = 5–16; 11 boys, 18 girls, 1 other) and their parents or caregivers (no metadata available) voluntarily participated in the study. Both the children and the adults were requested to add emoji magnets of their choosing to the same seven pre-constructed *WhatsApp* messages. This methodology was chosen because the data had to be collected in a task that was fun, uncomplicated, and suitable for the young children who would attend the language festival.

Fifty emoji magnets, with 49 different emoji, were available to the participants. These represented the following emoji: 😃😂😭😉😘😍😛😐😎😇😴🤓😺😹😻🙈🙉🐵👍👌🙏👊✌🙌💃👯🍔🍕🌮🎂🍆🍌🍻☕☕💕💔💩💀👻🌴🌸☀🔥🚗✈🚀💣👀💯. This was a standard emoji set, i.e., a convenience sample, that had been purchased by the researchers for the purposes of the data collection. One magnet set was available to the children and one magnet set to the adults. Therefore, participants could use each magnet/emoji only once and could not repeat emoji (except for the coffee emoji, which happened to occur twice in the magnet set).

The following seven messages in Dutch had been devised by the principal researcher for the addition of emoji (English translation provided below):

(1) *Yesss Morgen naar de Efteling voor mn verjaardag!!*
'Yesss Tomorrow to the Efteling for my birthday!!'

(2) *RIP! Kat Poekie van oma is overleden*
'RIP! Grandma's cat Poekie has passed away'

(3) *Lekker chillen op het strand #vakantie #genieten*
'Chilling on the beach #holiday #enjoy'

(4) *Whaha wat n blunder… In de poep gestapt, oeps!*
'Whaha what a blunder... Stepped in poo, oops!'

(5) *Zaterdag mogen we kiezen wat we eten. Jippieee*
'Saturday we can choose what we eat. Yaaay'

(6) *Grapje! Ik speel toch NOOOIT vals*
'Just kidding! I NEEEVER cheat anyway'

(7) *Hey sorry dat ik boos was … love you*
'Hey sorry that I was angry ... love you'

The messages were devised so as to match the range of emoji available in the magnet sets and aimed to resemble actual Dutch youths' *WhatsApp* messages. They were written to be suitable for primary school-aged children and were checked by two teachers for their appropriateness in terms of both language and content. As for language, the messages were intentionally informal and included features of textese, such as reduplications (*yesss*, *jippieee*, *NOOOIT*), hashtags (*#vakantie*, *#genieten*), interjections (*yesss*, *whaha*, *oeps*, *jippieee*, *hey*), non-standard abbreviations and orthography (*RIP*, *mn*, *n*, *hey*), and English borrowings (*yesss*, *RIP*, *chillen*, *love you*). In terms of content, they covered the topics of a birthday trip to a well-known Dutch amusement park, the passing away of a pet, a vacation, an unfortunate incident with poo, choosing dinner, cheating at games, and an apology for being angry. Three messages were happy in sentiment (1, 3, 5), three expressed more complex emotions (4, 6, 7), and one was clearly sad (2). The messages were visualised as *WhatsApp* chats and printed on large posters. As for the lay-out, extra spacing was provided at the beginning, in the middle, and at the end of each message, so as to leave room for positioning emoji anywhere. Participants were also allowed to add emoji right next to or below words, indicating that they should be inserted right after a word.

The poster with the *WhatsApp* messages was attached to a magnetic board. Each child and adult were positioned back to back, so the simultaneous data collection occurred independently (they could not see each other's emoji choices). Participants were instructed to decide for themselves how many emoji to use, which emoji to use, and where to add them. Afterwards, metadata on the child participants' gender, age, smartphone

ownership, and social media use[1] were gathered. Moreover, informed consent was collected of all participants.

Because the data collection involved underage children, we sought ethical approval beforehand. The data collection procedure was approved by *Radboud University's Ethics Assessment Committee*.

The Appendix presents a picture of what the collected data looked like. After a child and adult had added the emoji magnets to the *WhatsApp* messages, a picture of the poster with the messages and emoji was taken. In the end, this provided us with 60 pictures: 30 of emoji use by children and 30 of emoji use by adults. The next step was to digitise all the data: for each participant, the messages and emoji were copied in digital format into *Microsoft Excel*, including their exact use of emoji (as had been captured in the pictures). The corpus for investigating children's and adults' emoji preferences contained 420 messages (60 participants × 7 messages), with a total of 1012 emoji.

3.2. Procedure: Data coding

The corpus was coded in *Excel* for: a) the number of emoji per message, b) the type of each emoji that was used, c) the position of each emoji in the messages, d) the function of each emoji, and e) the sentiment conveyed by emoji that expressed sentiment.

Based on the emoji included in the magnet set used for collecting the data, a distinction was made between six types of emoji:

1) Faces (12): 😃😂😭😉😘😍😛😐😎😇😴🤓
2) Animal faces (6): 🐱😹🙀🙈🐵🐵
3) Gestures and movements (8): 👍👌🙏👊✌️👐💃👫
4) Food and drinks (8): 🍔🍕🌮🎂🍆🍌🍻☕
5) Hearts (2): 💕💔
6) Other (including objects and symbols) (13): 💩💀👻🌴🌸☀️🔥🚗✈️🚀💣👀💯

[1] Note that smartphone ownership and social media use did not correspond one to one, since children who did not own a smartphone could use their parents' phone for social media apps. In fact, all children reported having at least some experience with using a smartphone.

For the position of emoji, the coding scheme distinguished between four options of where to add the emoji to the pre-constructed messages: a) at the beginning of a message (before the text); b) after a keyword within the message; c) between sentences, clauses, or intonation units (in the middle of a message); or d) at the end of a message (after the text). These four positions are visualised in example (8):

(8) 😘 *Hey sorry dat ik boos* 🔥 *was ...* 🐵 *... love you* 💕
'😘 Hey sorry that I was angry 🔥 ... 🐵 ... love you 💕'

From an initial exploration of our corpus, four functions of emoji emerged: a) visualising a keyword in the message, b) visualising the content of a message, c) expressing an emotion, and d) unconventional use. The main distinction between the visualisation functions is that the emoji either literally matched a specific (key)word in a message (e.g., a palm tree emoji 🌴 accompanying the word *beach*; food emoji 🍕🌮🍔 accompanying the word *food*) or was associated by participants with the general content of a message but did not match any specific word (e.g., a plane ✈️ or a beer emoji 🍻 in a message about a holiday that made no mention of the travel mode or drinking of any kind). In example (8) above, the fire emoji visualises a keyword (*boos* 'angry') and the two hearts emoji expresses emotion. In example (9) below, the birthday cake emoji also visualises a keyword (*verjaardag* 'birthday'), while the car emoji visualises the general message, but not a specific word within the message. Emoji use was coded as 'unconventional' when it did not correspond to any of the conventional meanings of the emoji as codified by Emojipedia and when it otherwise made no sense to the annotator: for instance, when an adult participant used the 'face with tears of joy' in a message that expressed a sad occasion, such as the death of a pet in example (10). Emoji could also be coded for multiple functions (but this was the case for only 3,5% of all emoji in the corpus).

(9) *Yesss Morgen naar de Efteling voor mn verjaardag!!* 🚗😉🎂
'Yesss Tomorrow to the Efteling for my birthday!! 🚗😉🎂'

(10) *RIP!* 😿 *Kat Poekie van oma is overleden* 💀😂
'RIP! 😿 Grandma's cat Poekie has passed away 💀😂'

For the emoji whose function was to convey a sentiment, the sentiment of emoji was specified in coding the data. Our coding scheme made a distinction between positive sentiment (expressing happiness, amusement, joy, or love, e.g., 😃🌸😂👍🏻😍), negative sentiment (expressing sadness, anger, or fear, e.g., 😭💔💣😕😿), and ambiguous

sentiment (open to multiple interpretations (subjective), e.g., 🙈👀🤓). Note that since emoji can express subtle emotions and a broad spectrum of sentiments (Novak *et al.* 2015; Upadhyay *et al.* 2023), this classification is an oversimplification, but the positive-negative dichotomy is at the core of much research on emotions (Solomon and Stone 2002) and has been used in recent emoji research (e.g. Neel *et al.* 2023). Emoji were classified by the annotator on a case-by-case basis in the context of the message in which they were used.

The codebook was established by scrutinizing a subset of the data. After practising with the codebook, the entire corpus was coded independently by the second author. When in doubt, specific cases were discussed with the first author, until a consensus was reached.

3.3. Statistical treatment: Data analysis

The statistical analysis of the coded data consisted of two parts. The first research question set out to compare and contrast children's and adults' use of emoji. A t-test was performed to examine if there was an effect of age group (children vs. adults) on the number of emoji that were used per message. Chi-square tests were conducted to examine if there were relationships between age group and the type, function, or sentiment of the emoji used. A Fisher's exact test was run for testing if there was a relationship between age group and the position of emoji, since not all requirements for a chi-square test (i.e., at least five observations per condition) were met.

The second research question aimed to identify factors that affect children's use of emoji. To answer this question, children's emoji use was analysed together with their Age (5 to 7 years old, 8 to 9 years old, 10 to 16 years old),[2] Gender (girls vs. boys), smartphone ownership (yes vs. no), and social media use (yes vs. no). First, Pearson correlations were calculated between these four variables. There turned out to be significant correlations between age and social media use ($r(28) = .451$, $p = .014$) and between gender and smartphone ownership ($r(28) = .391$, $p = .033$). A closer inspection showed that older

[2] Age was divided into three groups with a similar number of child participants, for performing chi-square tests.

children more frequently used social media than younger children and that girls more often possessed their own smartphone than boys.

Then, a multiple linear regression was conducted to explore if any of these variables (children's age, gender, owning a smartphone, and using social media) predicted the number of emoji per message. Since age and social media use correlated significantly, as well as gender and smartphone ownership, only age and smartphone ownership were added as predictor variables in the regression model. These two were selected because another (exploratory) regression suggested that they would contribute more to the model than their correlating counterparts.

Finally, more chi-square tests were performed to investigate relationships between, on the one hand, the children's demographic variables (age, gender, smartphone ownership, and social media use) and, on the other hand, the type, position, function, and sentiment of emoji. When there were not enough observations in a condition to meet the requirements for chi-square testing, a Fisher's exact test was performed instead.

4. Results

4.1. Children's vs. adults' emoji use

4.1.1. Number of emoji

To explore if there was any effect of age group on the number of emoji used per message, a simple t-test was performed. No significant difference was found ($t(58) = -.06$, $p = .953$) between the number of emoji that children used ($M = 2.40$, $SD = .065$) and the number of emoji that adults used ($M = 2.41$, $SD = .585$) in our corpus. In fact, the number of emoji used in total by all of the child participants and the number of emoji used by all of the adult participants was nearly identical, 505 and 507 respectively.

4.1.2. Type, position, function, and sentiment of emoji

An overview of the raw frequencies of emoji use by participants of both age groups can be found in Tables 1–4 below. Because the total number of emoji used by the children and adults were so similar, it was deemed unnecessary to compute relative frequencies.

Age group	Faces	Animal faces	Gestures and movements	Food and drinks	Hearts	Other
Children	170	44	43	96	39	113
Adults	170	48	58	88	25	118

Table 1: Frequencies of type of emoji by age group

Age group	At beginning of message	After keyword (within message)	Between sentences (in the middle of message)	At end of message
Children	3	7	144	351
Adults	11	3	144	349

Table 2: Frequencies of position of emoji by age group

Age group	Visualisation of keyword	Visualisation of message	Expression of emotion	Unconventional use
Children	168	35	305	12
Adults	179	42	288	5

Table 3: Frequencies of function of emoji by age group

Age group	Positive sentiment	Negative sentiment	Ambiguous sentiment
Children	205	70	30
Adults	199	56	33

Table 4: Frequencies of sentiment of emoji expressing emotion by age group

Chi-square tests and a Fisher's exact test were carried out to investigate if there were any relationships between age group and the type, sentiment, or function of the emoji used. As seen in Table 5, no significant differences were found between the children and adults.

Type	Purpose	Sentiment	Position
$\chi^2 = 5.92$	$\chi^2 = 4.32$	$\chi^2 = 1.30$	
$p = .314$	$p = .229$	$p = .522$	$p = .102$ (F)

Table 5: Results of χ^2 tests and Fisher's exact tests[3]

[3] *Note*: (F) = Fisher's exact test was performed instead of χ^2 test

4.2. Children's emoji use

4.2.1. Number of emoji

A multiple linear regression was performed to investigate which variables predicted the number of emoji per message. Because of correlations between children's age and social media use and between gender and owning a smartphone, social media use and gender were excluded and only age and smartphone ownership were included in the regression.

The regression showed that there was a collective significant effect of age and smartphone ownership on the number of emoji used ($F(2,26) = 5.07$, $p = .014$, $R^2 = .281$). The individual predictors were examined further and indicated that only age was a significant predictor in the model ($\beta = .521$, $p = .007$). Closer inspection of the data showed that the older the children were, the more emoji they used per message. Means and standard deviations for the three age groups that we distinguished among the children are shown in Table 6.

5 to 7 years old	**8 to 9 years old**	**10 to 16 years old**
$M = 1.91$	$M = 2.60$	$M = 2.77$
$SD = 0.54$	$SD = 0.71$	$SD = 0.32$

Table 6: Means and standard deviations per age group

4.2.2. Type, position, function, and sentiment of emoji

Chi-square tests were performed to investigate the relationships between the independent variables (age, gender, smartphone ownership, social media use) and the dependent variables (type, position, function, and sentiment of emoji). In Table 7 below we can see that no significant relationships were found between the independent variables and the type of emoji or sentiment of emoji that were used. Significant relationships between the independent variables and the function of emoji and the position of emoji are reported in more detail below.

Independent variable		Type	Function	Sentiment	Position
Age	(5–7 vs. 8–9 years)	$\chi^2 = 5.30$	$\chi^2 = 6.39$	$\chi^2 = 0.12$	
		$p = .381$	$p = .094$	$p = .943$	$p < .001$*** (F)
	(8–9 vs. 10–16 years)	$\chi^2 = 1.39$		$\chi^2 = 2.40$	
		$p = .926$	$p = .051$ (F)	$p = .302$	$p = .096$ (F)
	(5–7 vs. 10–16 years)	$\chi^2 = 3.34$		$\chi^2 = 1.36$	
		$p = .647$	$p = .009$** (F)	$p = .507$	$p < .001$*** (F)
Gender (girl/boy)		$\chi^2 = 2.62$		$\chi^2 = 1.64$	
		$p = .758$	$p = .011$* (F)	$p = .440$	$p = .010$** (F)
Owns a smartphone (yes/no)		$\chi^2 = 3.80$	$\chi^2 = 7.19$	$\chi^2 = 1.59$	
		$p = .579$	$p = .066$	$p = .452$	$p < .001$*** (F)
Uses social media (yes/no)		$\chi^2 = 5.30$		$\chi^2 = 0.33$	
		$p = .381$	$p = .009$** (F)	$p = .846$	$p < .001$*** (F)

Table 7: Results of χ^2 tests and Fisher's exact tests

4.2.2.1 Position of emoji and age

There was a significant relationship between children's age and the position in a message where they inserted the emoji. This difference was visible between the group of 5-to-7-year-old participants compared to the group of 8-to-9-year-old participants ($p < .001$) and between the group of 5-to-7-year-old participants compared to the group of 10-to-16-year-old participants ($p < .001$). Standardised residuals, provided in Table 8, show that 5-to-7-year-olds more often put their emoji at the end of a message than 8-to-9-year-olds and 10-to-16-year-olds. Furthermore, 5-to-7-year-olds less often placed their emoji after a sentence in the middle of a message than 8-to-9-year-olds and 10-to-16-year-olds.

Group	At beginning of message	After keyword (within message)	Between sentences (in the middle of message)	At end of message
5–7 years	NaN	-1.647	-3.152	3.515
8–9 years	NaN	1.647	3.152	-3.515
5–7 years	-1.619	-1.619	-4.245	4.874
10–16 years	1.619	1.619	4.245	-4.874

Table 8: Standardised residuals of children's age and the position of emoji

4.2.2.2 Position of emoji and gender

There was a significant relationship between gender and where in a message emoji were most often positioned ($p < .01$). Standardised residuals are presented in Table 9. They show that girls less often put emoji at the end of a message and more often placed them after a sentence in the middle of a message than boys did.

Group	At beginning of message	After keyword (within message)	Between sentences (in the middle of message)	At end of message
Girls	1.298	1.840	2.177	-2.800
Boys	-1.298	-1.840	-2.177	2.800

Table 9: Standardised residuals of children's gender and the position of emoji

4.2.2.3 Position of emoji and smartphone ownership

There was a significant relationship between owning a smartphone and the most frequent positioning of emoji in a message ($p < .001$). Standardised residuals, as shown in Table 10, show that children who owned a smartphone placed emoji at the end of a message less often. They rather placed them after a sentence in the middle of a message. This was not the case for children who did not have their own smartphone.

Group	At beginning of message	After keyword (within message)	Between sentences (in the middle of message)	At end of message
Does not own a smartphone	-0.922	-1.414	-3.487	3.931
Owns a smartphone	0.922	1.414	3.487	-3.931

Table 10. Standardised residuals of children's smartphone ownership and the position of emoji

4.2.2.4 Position of emoji and social media use

There was a significant relationship between using social media and the most frequent positioning of emoji in a message ($p < .001$). Table 11 presents the standardised residuals. The data show that children who used social media less often placed emoji at the end of a message, and that they also placed emoji more in the middle of a message than children who were not used to social media.

Group	At beginning of message	After keyword (within message)	Between sentences (in the middle of message)	At end of message
No social media	-0.243	-0.632	-4.399	4.515
Social media use	0.243	0.632	4.399	-4.515

Table 11: Standardised residuals of children's social media use and the position of emoji

4.2.2.5 Function of emoji and age

There was a significant relationship between age and the functions of the emoji used. This relationship was visible in the difference between the group of 5-to-8-year-old participants compared to the group of 10-to-16-year-old participants ($p < .01$). Standardised residuals (see Table 12), show that the emoji used by 5-to-7-year-olds were less often meant to visualise a keyword than those used by 10-to-16-year-olds. Furthermore, the emoji use of 5-to-7-year-olds was more often unconventional than the emoji use of 10-to-16-year-olds, who did not use emoji in an unconventional manner.

Group	Visualisation of keyword	Visualisation of message	Expression of emotion	Unconventional use
5–7 years	-2.193	0.562	1.053	2.666
10–16 years	2.193	-0.562	-1.053	-2.666

Table 12: Standardised residuals of children's age and the function of emoji

4.2.2.6 Function of emoji and gender

There was a significant relationship between gender and the functions of the emoji used ($p < .05$). Table 13 provides the standardised residuals, which show that girls more often used emoji to visualise a keyword than boys. In addition, boys' emoji use was more unconventional than that of girls.

Group	Visualisation of keyword	Visualisation of message	Expression of emotion	Unconventional use
Girls	2.670	-0.438	-1.610	-2.253
Boys	-2.670	0.438	1.610	2.253

Table 13: Standardised residuals of children's gender and the function of emoji

4.2.2.7 Function of emoji and social media use

Finally, there was a significant relationship between using social media and the functions of the emoji used *($p < .01$)*. Standardised residuals (Table 14) show that children who did not use social media used emoji more often in an unconventional way than children who reported using social media.

Group	Visualisation of keyword	Visualisation of message	Expression of emotion	Unconventional use
No social media	-1.761	0	0.727	3.100
Social media use	1.761	0	-0.727	-3.100

Table 14: Standardised residuals of children's social media use and the function of emoji

5. DISCUSSION

This paper reports on a corpus study that set out to explore children's emoji use, which has remained understudied in previous research. Prior studies into emoji and children focused mostly on the potential of emoji to express emotions, including children's perceptions of facial emoji (Oleszkiewicz *et al.* 2017; Liu and Li 2021), how emoji can help children to understand emotions and concepts (Fane 2017; Fane *et al.* 2018) and how emoji can measure children's emotions and preferences (Gallo *et al.* 2017; Schouteten *et al.* 2018; Swaney-Stueve *et al.* 2018; Lima *et al.* 2019), and the effects of emoji on children from a marketing perspective (Siegel *et al.* 2015; Luangrath *et al.* 2017). For the present study, a corpus was collected in an innovative manner for the sole purpose of eliciting emoji use from children and their parents or caregivers. A pragmalinguistic approach was taken to analyse the corpus.

5.1. A comparison of children's and adults' use of emoji

Our corpus analysis started with a comparison between children's and adults' emoji use. The first part of our results showed no significant differences between children and adults in their use of emoji. Both age groups preferred facial emoji (in favour of other types, such as objects), placed emoji mostly at the end of messages (rather than at the beginning, after a keyword, or between sentences, clauses, or intonation units), used emoji to express emotions mostly (instead of visualising keywords or the content of a message), and

mostly used emotional emoji to convey a positive sentiment, rather than a negative or ambiguous one.[4] Realtime monitoring of emoji use on *Twitter* with emojitracker[5] shows that the top ten of most popular emoji by adults on *Twitter* include many faces (😂😭😍😊😔😘) and hearts (❤️❤️💕), which also express emotions. Likewise, the top ten of the most frequently used emoji in a large Swiss *WhatsApp* corpus only consisted of (emotional) faces (😂😄😊😉😘😁☺️), animal faces (🙈😻), and a heart (❤️). In fact, 65 per cent of all emoji in the corpus were facial emoji (Dürscheid and Siever 2017). Prior research based on large-scale *Twitter* corpora showed that the most popular emoji have a positive sentiment and that adults tend to use emoji at the end of tweets (Novak *et al.* 2015). Similarly, the most popular emoticon :-) is positive in sentiment and the majority of emoticons (64%) appear at the end of tweets (Spina 2018). The current study, albeit with a small-scale corpus analysis, confirms all these findings on emoji use and tentatively extends them to children as well.

Our results suggest that adults and children generally have very similar intuitions about how many emoji to add to a message, which emoji to use (of which types), and where to place them. Furthermore, the emoji that were used by adults and children often served similar pragmatic functions in the message and overall held similar sentiments. The latter finding concurs with results from previous studies on children and emotions: like adults, children can attribute emotions to commonly used emoji (Oleszkiewicz *et al.* 2017; Liu and Li 2021; da Quinta *et al.* 2023) and can therefore select emoji that match the sentiment of a message.

The findings mentioned above answer our first research question: the study has not provided evidence that children (digital natives) use emoji differently than adults (digital immigrants). If all conditions are equal —i.e., when presented with the same 'digital' messages and the same set of emoji to choose from— digital natives and digital immigrants do not seem to use emoji in significantly different ways. This appears not to be in line with Frey and Glaznieks's (2018) finding that digital natives use more CMC style markers, which include emoji, than digital immigrants. This discrepancy may be due to their methodological choice of not separating emoji from other markers such as

[4] The sentiment that was visually expressed by the emoji being used matched the sentiment that was verbally expressed in the pre-constructed messages, since there were more messages with positive than negative content.
[5] http://emojitracker.com

emoticons, acronyms, and hashtags. It may also be because of an age difference between the digital natives who contributed to their corpus (operationalised quite broadly as people born from 1980 onwards) and the digital natives who contributed to our corpus (children born between 2004 and 2015 who are quite young in comparison). Still, our study shows that both digital immigrants and children who are digital natives reveal a basic and similar emoji literacy (Danesi 2016; Scheffler *et al.* 2022).

5.2. A closer inspection of children's use of emoji

The study's second research question aimed to identify which demographic factors affect children's use of emoji. Our results showed that the older the children, the more emoji they tended to use per message. The connection between emoji literacy and traditional literacy (Scheffler *et al.* 2022) could be at play here. The older children may have had a better understanding of the pre-constructed social media messages than the younger children, and therefore a better understanding of the added value of emoji to the text. This effect might also be attributed to the significant positive correlation between age and social media use. In other words, the greater exposure to social media of older children as compared to younger children is another possible explanation for their greater use of emoji. Emoji use keeps increasing and about one in five tweets now contains at least one emoji (Emojipedia)[6] and, for *WhatsApp*, this number is even higher, as Dürscheid and Siever (2017) report that a staggering 91 percent of all *WhatsApp* chats in their corpus contained emoji. Accordingly, more exposure to social media will be inextricably linked to more exposure to emoji.

Several significant relationships were found for the function and the position of emoji but not for the type or for the sentiment of emoji. It is conceivable that this lack of differences for the variables of type and sentiment is due to our method of corpus collection. All participants added emoji to the same messages, the content of which guided the children in the types of emoji they selected. For instance, the message about what to choose for dinner caused the selection of food emoji. Likewise, the content of the messages as well as the emoji that were available in the magnet set invited children to use emoji with a certain sentiment. The message about a birthday, for example, elicited emoji with a positive sentiment, whereas the message about the death of a pet could be expected

[6] https://emojipedia.org

to elicit emoji with a negative sentiment. Regarding the position and function of the emoji, the participants were not guided in any way, making these significant relationships between children's demographic variables and emoji use variables especially interesting.

Children's age, gender, social media use, and smartphone ownership were in significant relationship with the position of emoji. Being older, being a girl, owning a smartphone, and using social media were all related to positioning fewer emoji at the end of messages and more emoji in the middle of messages, between sentences. A possible explanation for this may lie in the established use of emoji instead of final punctuation marks to end sentences in written CMC (Danesi 2016; Sampietro 2016). The end of a message can be considered the default position for emoji: both adults and teenagers tend to conclude their messages with one or more emoji systematically (Novak *et al.* 2015; Hilte *et al.* 2022). While the use of a period has become pragmatic rather than syntactic in informal digital writing (Androutsopoulos and Busch 2021), emoji, in contrast, have assumed a structural or syntactic role, similar to emoticons (Provine *et al.* 2007; Dresner and Herring 2010; Spina 2018), replacing traditional punctuation marks (Dürscheid and Siever 2017; Pappert 2017; Beißwenger and Pappert 2019; Busch 2021). Placing emoji between sentences, clauses, or intonation units within a message as structural boundaries can thus be seen a more sophisticated use of emoji. Both owning a smartphone and using social media expose children to this use of emoji as sentence boundaries, and older children will have received more such exposure, making smartphone owners, social media users, and older children more emoji literate and thus more aware of the possibility to use emoji in a punctuation-like manner. The gender difference here may partly depend on the correlation between gender and smartphone ownership, where girls possessed their own smartphone more often than boys.

Age, gender, and social media use were also significantly related to the functions of the emoji used in the message. Firstly, participants who were older, female, or were social media users used emoji in an unconventional manner less frequently. Assuming that using social media increases exposure to emoji and hence emoji literacy, this last relationship makes sense. The finding that girls and older children used emoji less in unconventional ways may be explained by their traditional (linguistic) literacy skills, which are attested to be higher in these demographic groups than among boys and younger children (Below *et al.* 2010; McTigue *et al.* 2021). These findings also concur with those of Oleszkiewicz *et al.* (2017), who found that older children and girls are more

adept at recognizing the emotions that are expressed with emoji. Secondly, the girls and older children more often used emoji to visualise a keyword. Perhaps there is, again, a connection with traditional literacy. Better literacy skills could cause participants to pay more attention to the individual words in a text, using emoji to visualise them and thus disambiguate the content of a message (Riordan 2017b; Beißwenger and Pappert 2019).

5.3. Limitations and future research

The exploratory nature of the present corpus study has some drawbacks. Because our corpus was collected in a semi-experimental setting, this analysis should be replicated with a corpus of natural social media messages to examine if similar patterns occur. The manner in which the data were obtained also entailed that the interlocutor (to whom the social media messages were hypothetically sent) was unclear, which means that we could not explore children's social use of emoji. In a non-experimental CMC setting, people with high emoji literacy may be more mindful of their conversation partner and the conversational setting in their emoji usage, showing more situational awareness. Following Hilte *et al.* (2021), further research could discover if people of different age groups accommodate to their interlocutor to a greater or lesser extent in their emoji use. Wang *et al.*'s (2019) distinction between communicative and performative use of emoji (and other graphicons) also deserves more attention in future research: when do people use emoji for instrumental purposes, to support communication, and when do they merely want to show off their emoji literacy to their audience?

The ecological validity of our study was subject to limitations. The 50 emoji magnets that participants could use during data collection represent a very small subset of the current range of over 3,700 existing emoji (Emojipedia)[7] and were not a representative sample in terms of types: for example, no activities (🚵, 🏄) were included. Although participants had popular emoji at their disposal, future studies would preferably not limit participants in their emoji selection. Since it is unfeasible to select from thousands of emoji magnets, a recommendation for future studies is to collect data in a digital fashion, which would enable participants to a) select from all existing emoji, b) use the same emoji more than once, and c) find emoji using keywords.

---

[7] https://emojipedia.org/stats

In our analyses, the adult participants were treated as a homogeneous group, because the present study's main interest lies with children's use of emoji. Metadata about the adults, such as demographic information and social media use, were unfortunately not collected. Among the child participants, differences could be identified regarding their age and use of social media. A closer inspection indicated that these variables impacted children's emoji use in multiple ways. This raises intriguing questions regarding the nature of emoji literacy. Can differences in emoji use or understanding between children and adults be found if children are somewhat older (i.e., pre-teens or teenagers) and more experienced social media users? Future research could delve into the question if digital immigrants use emoji much in the same way as young children who are not yet experienced social media users. It is plausible that more advanced emoji literacy — including extensive knowledge of connotations of emoji (Weissman 2019), of how emoji can have different literal and figurative meanings depending on the context— only emerges in (pre)adolescence, when children have gained more experience with emoji and have had more exposure to other users' emoji. In other words, emoji literacy is likely to go hand in hand with familiarity with emoji.

Since the participating children's literacy skills were not tested, we can only speculate how age effects can be explained. They may even be due to differences in reading skills or in properly understanding the task at hand. The youngest participant was only five years old: her limited reading proficiency may have hindered her understanding of the messages and, consequently, her execution of the task. Further research should be undertaken to determine the interplay between traditional literacy and emoji literacy.

Finally, the corpus was rather small: 60 participants contributed to it. Its scale was limited by the analogue data collection at a language festival. Children's emoji use could be investigated with a larger corpus, to allow for more external validity. However, our corpus did contain over a thousand emoji, which justifies our quantitative statistical analyses. Because all participants were Dutch, results may not be generalisable cross-linguistically or cross-culturally, since emoji use has been shown to differ across cultures (Barbieri *et al.* 2016; Freedman 2018; Guntuku *et al.* 2019). This invites future research to take a contrastive cross-cultural perspective to children's emoji use.

5.4. Implications and conclusion

Although there is abundant room for further progress in exploring children's use of graphicons in written CMC, the findings of this study, while preliminary, constitute an important first step to the study of emoji use by (Dutch) children. Moreover, it has brought us a little closer to developing a full picture of the concept of emoji literacy. Besides such scientific relevance, our results may also have potential societal relevance for education, health communication, and marketing. First, wider knowledge about children's emoji use may be beneficial for educators as they teach children to read and write, as well as for the purposes of second language acquisition. Second, it may be valuable for child psychologists and paediatricians in doctor-patient conversations when trying to connect with this young age group. Lastly, it may help marketeers to further tailor their messages to a young target audience.

Our corpus study suggests that under identical circumstances, children and adults do not significantly differ in their use of emoji. What is more, we have shown that children's emoji use can be explained in terms of age, gender, smartphone ownership, and social media use. Being an older child, a girl, owning a smartphone, and using social media apps are all features related to more sophisticated emoji use and more emoji literacy.

REFERENCES

Abril, Danielle. 2022. *Gen Z Came to 'Slay.' Their Bosses don't Know what that Means*. *The Washington Post*. https://www.washingtonpost.com/technology/2022/12/12/gen-z-work-emojis

Androutsopoulos, Jannis and Florian Busch. 2021. Digital punctuation as an interactional resource: The message-final period among German adolescents. *Linguistics and Education* 62: 100871. https://doi.org/10.1016/j.linged.2020.100871

Austin, John L. 1962. *How to Do Things with Words*. Oxford: Clarendon.

Bai, Qiyu, Qi Dan, Zhe Mu and Maokun Yang. 2019. A systematic review of emoji: Current research and future perspectives. *Frontiers in Psychology* 10: 2221, 1–16. https://doi.org/10.3389/fpsyg.2019.02221

Barbieri, Francesco, German Kruszewski, Francesco Ronzano and Horacio Saggion. 2016. How cosmopolitan are emojis? Exploring emojis usage and meaning over different languages with distributional semantics. In Luca Rossetto, Ivan Giangreco, Claudiu Tanase and Heiko Schuldt eds. *Proceedings of the 24th ACM International Conference on Multimedia*. New York: Association for Computing Machinery, 531–535.

Beißwenger, Michael and Steffen Pappert. 2019. *Handeln mit Emojis: Grundriss einer Linguistik kleiner Bildzeichen in der WhatsApp-Kommunikation*. Duisburg-Essen: Universitätsverslag Rhein-Ruhr.

Below, Jaime L., Christopher H. Skinner, Jamie Y. Fearrington and Christy A. Sorrell. 2010. Gender differences in early literacy: Analysis of kindergarten through fifth-grade dynamic indicators of basic early literacy skills probes. *School Psychology Review* 39/2: 240–257.

Busch, Florian. 2021. The interactional principle in digital punctuation. *Discourse, Context & Media* 40: 100481. https://doi.org/10.1016/j.dcm.2021.100481

Cohn, Neil, Jan Engelen and Joost Schilperoord. 2019. The grammar of emoji? Constraints on communicative pictorial sequencing. *Cognitive Research: Principles and Implications* 4/33: 1–18.

Cohn, Neil, Tim Roijackers, Robin Schaap and Jan Engelen. 2018. Are emoji a poor substitute for words? Sentence processing with emoji substitutions. In Chuck Kalish, Martina A. Rau, Xiaojin (Jerry) Zhu and Timothy T. Rogers eds. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Seattle: Cognitive Science Society, 1524–1529.

Coosto. 2020. *Nationaal Emoji Onderzoek 2020*. https://www.coosto.com/nl/blogs/het-nationaal-emoji-onderzoek-2020

Danesi, Marcel. 2016. *The Semiotics of Emoji: The Rise of Visual Language in the Age of the Internet*. London: Bloomsbury.

Da Cruz, Marina F., Ramon S. Rocha, Ramon Silva, Mônica Q. Freitas, Tatiana C. Pimentel, Erick A. Esmerino, Adriano G. Cruz, Tatiana K. da S. Fidalgo and Lucianne C. Maia. 2021. Probiotic fermented milks: Children's emotional responses using a product-specific emoji list. *Food Research International* 143: 110269. https://doi.org/10.1016/j.foodres.2021.110269

Da Quinta, Noelia, Elena Santa Cruz, Yolanda Ríos, Begoña Alfaro and Íñigo Martínez de Marañón. 2023. What is behind a facial emoji? The effects of context, age, and gender on children's understanding of emoji. *Food Quality and Preference* 105/3: 104761. http://dx.doi.org/10.1016/j.foodqual.2022.104761

Dainas, Ashley R. and Susan C. Herring. 2021. Interpreting emoji pragmatics. In Chaoqun Xie, Francisco Yus and Hartmut Haberland eds. *Approaches to Internet Pragmatics: Theory and Practice*. Amsterdam: John Benjamins, 107–144.

De la Rosa-Carrillo and Ernesto L. 2018. Emoji literacies: Read & write, translate, montage. In Danilo M. Baylen ed. *Senses and Experiences: The Book of Selected Readings*. International Visual Literacy Association, 17–30. https://issuu.com/ivla.tbsr/docs/2018_tbsr

Deubler, Grace, Marianne Swaney-Stueve, Tegan Jepsen and Belinda P. Su-Fern. 2020. The K-State emoji scale. *Journal of Sensory Studies* 35/1: e12545, 1–9. https://doi.org/10.1111/joss.12545

Dijkmans, Corné, Peter Kerkhof and Camiel Beukeboom. 2020. Adapting to an emerging social media landscape: The rise of informalization of company communication in tourism. In Julia Neidhardt and Wolfgang Wörndl eds. *Information and Communication Technologies in Tourism 2020*. Cham: Springer, 3–14.

Dresner, Eli and Susan C. Herring. 2010. Functions of the nonverbal in CMC: Emoticons and illocutionary force. *Communication Theory* 20/3: 249–268.

Dürscheid, Christa and Dimitrios Meletis. 2019. Emoji: a grapholinguistic approach. In Yannis Haralambous ed. *Proceedings of Graphemics in the 21st Century*. Brest: Fluxus Editions, 167–183.

Dürscheid, Christa and Christina M. Siever. 2017. Jenseits des alphabets: Kommunikation mit emoji. *Zeitschrift für Germanistische Linguistik* 45/2: 256–285.

EditieNL. 2016. *Appen Met pa en ma: Waarom Gaat het Toch zo vaak Mis? RTL Nieuws*. https://www.rtlnieuws.nl/editienl/artikel/606736/appen-met-pa-en-ma-waarom-gaat-het-toch-zo-vaak-mis

Evans, Vyvyan. 2017. *The Emoji Code: How Smiley Faces, Love Hearts and Thumbs Up are Changing the Way we Communicate*. London: Michael O'Mara.

Fane, Jennifer. 2017. Using emoji as a tool to support children's well-being from a strength-based approach. *Learning Communities Journal* 21: 96–107.

Fane, Jennifer, Colin MacDougall, Jessie Jovanovic, Gerry Redmond and Lisa Gibbs. 2018. Exploring the use of emoji as a visual research method for eliciting young children's voices in childhood research. *Early Child Development and Care* 188/3: 359–374.

Franco, Courtny L. and Jennifer M.B. Fugate. 2020. Emoji face renderings: Exploring the role emoji platform differences have on emotional interpretation. *Journal of Nonverbal Behavior* 44/2: 301–328.

Freedman, Alisa. 2018. Cultural literacy in the empire of emoji signs: Who is 😂? In Elena Giannoulis and Lukas R.A. Wilde eds. *Emoticons, Kaomoji, and Emoji: The Transformation of Communication in the Digital Age*. New York: Routledge, 44–66.

Frey, Jennifer-Carmen and Aivars Glaznieks. 2018. The myth of the digital native? Analysing language use of different generations on Facebook. In Reinhild Vandekerckhove, Darja Fišer and Lisa Hilte eds. *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora*. Antwerp: University of Antwerp, 41–44.

Gallo, Katherine E, Marianne Swaney-Stueve and Delores H. Chambers. 2017. A focus group approach to understanding food-related emotions with children using words and emojis. *Journal of Sensory Studies* 32/3: e12264. https://doi.org/10.1111/joss.12264

Gawne, Lauren and Gretchen McCulloch. 2019. Emoji as digital gestures. *Language@Internet* 17: https://web.archive.org/web/20240127230708/https://www.languageatinternet.org/articles/2019/gawne

Guntuku, Sharath C., Mingyang Li, Louis Tay and Lyle H. Ungar. 2019. Studying cultural differences in emoji usage across the East and the West. In Jürgen Pfeffer ed. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. Munich: Association for the Advancement of Artificial Intelligence, 226–235.

Herring, Susan C. and Ashley R. Dainas. 2017. "Nice picture comment!" Graphicons in Facebook comment threads. In Tung X. Bui and Ralph Jr. Sprague eds. *Proceedings of the 50th Hawaii International Conference on System Sciences*. Manoa: University of Hawaii at Manoa, 2185–2194.

Herring, Susan C. and Ashley R. Dainas. 2020. Gender and age influences on interpretation of emoji functions. *ACM Transactions on Social Computing* 3/2: 1–26.

Hilte, Lisa, Walter Daelemans and Reinhild Vandekerckhove. 2021. Interlocutors' age impacts teenagers' online writing style: Accommodation in intra-and intergenerational online conversations. *Frontiers in Artificial Intelligence* 4: 738278. https://doi.org/10.3389/frai.2021.738278

Hilte, Lisa, Reinhild Vandekerckhove and Walter Daelemans. 2022. Linguistic accommodation in teenagers' social media writing: Convergence patterns in mixed-gender conversations. *Journal of Quantitative Linguistics* 29/2: 241–268.

Hurlburt, George. 2018. Emoji: Lingua franca or passing fancy? *IT Professional* 20/5: 14–19.

Lima, Mayara, Marcela de Alcantara, Inayara B.A. Martins, Gastón Ares and Rosires Deliza. 2019. Can front-of-pack nutrition labeling influence children's emotional associations with unhealthy food products? An experiment using emoji. *Food Research International* 120: 217–225.

Liu, Siying and Na Li. 2021. Going virtual in the early years: 30-month-old toddlers recognize commonly used emojis. *Infant Behavior and Development* 63: 101541. https://doi.org/10.1016/j.infbeh.2021.101541

Luangrath, Andrea W., Joann Peck and Victor A. Barger. 2017. Textual paralanguage and its implications for marketing communications. *Journal of Consumer Psychology* 27/1: 98–107.

Manganari, Emmanouela E. 2021. Emoji use in computer-mediated communication. *International Technology Management Review* 10/1: 1–11.

Massey, Simon. 2022. Using emojis and drawings in surveys to measure children's attitudes to mathematics. *International Journal of Social Research Methodology* 25/6: 877–889.

McTigue, Erin M., Knut Schwippert, Per H. Uppstad, Kjersti Lundetræ and Oddny J. Solheim. 2021. Gender differences in early literacy: Boys' response to formal instruction. *Journal of Educational Psychology* 113/4: 690–705.

Miller, Hannah, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In Derek Ruths ed. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence, 152–161.

Miller, Hannah, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen and Brent Hecht. 2016. "Blissfully happy" or "ready to fight": Varying interpretations of emoji. *Proceedings of the Tenth International AAAI Conference on Web and Social Media*. Quebec: Association for the Advancement of Artificial Intelligence, 259–268.

Neel, Louise A.G., Jacqui G. McKechnie, Christopher M. Robus and Christopher J. Hand. 2023. Emoji alter the perception of emotion in affectively neutral text messages. *Journal of Nonverbal Behavior* 47/1: 83–97.

Novak, Petra K., Jasmina Smailović, Borut Sluban and Igor Mozetič. 2015. Sentiment of emojis. *PLOS ONE* 10/12: e0144296. https://doi.org/10.1371/journal.pone.0144296

Oleszkiewicz, Anna, Tomasz Frackowiak, Agnieszka Sorokowska and Piotr Sorokowski. 2017. Children can accurately recognize facial emotions from emoticons. *Computers in Human Behavior* 76: 372–377.

Pappert, Steffen. 2017. Zu kommunikativen funktionen von emojis in der whatsapp-kommunikation. In Michael Beißwenger ed. *Empirische Erforschung Internetbasierter Kommunikation*. Berlin: De Gruyter, 175–211.

Prada, Marília, David L. Rodrigues, Margarida V. Garrido, Diniz Lopes, Bernardo Cavalheiro and Rui Gaspar. 2018. Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics* 35/7: 1925–1934.

Prensky, Marc. 2001. Digital natives, digital immigrants. *On the Horizon* 9/5: 1–6.

Provine, Robert R., Robert J. Spencer and Darcy L. Mandell. 2007. Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology* 26/3: 299–307.

Riordan, Monica A. 2017a. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology* 36/5: 549–567.

Riordan, Monica A. 2017b. The communicative role of non-face emojis: Affect and disambiguation. *Computers in Human Behavior* 76: 75–86.

Sampietro, Agnese. 2016. Exploring the punctuating effect of emoji in Spanish WhatsApp chats. *Lenguas Modernas* 1/47: 91–113.

Scheffler, Tatjana, Lasse Brandt, Marie de la Fuente and Ivan Nenchev. 2022. The processing of emoji-word substitutions: A self-paced-reading study. *Computers in Human Behavior* 127: 107076. https://doi.org/10.1016/j.chb.2021.107076

Schouteten, Joachim J., Jan Verwaeren, Xavier Gellynck and Valérie L. Almli. 2019. Comparing a standardized to a product-specific emoji list for evaluating food products by children. *Food Quality and Preference* 72: 86–97.

Schouteten, Joachim J., Jan Verwaeren, Sofie Lagast, Xavier Gellynck and Hans De Steur. 2018. Emoji as a tool for measuring children's emotions when tasting food. *Food Quality and Preference* 68: 322–331.

Seargeant, Philip. 2019. *The Emoji Revolution: How Technology is Shaping the Future of Communication*. Cambridge: Cambridge University Press.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Sick, Julia, Erminio Monteleone, Lapo Pierguidi, Gastón Ares and Sara Spinelli. 2020a. The meaning of emoji to describe food experiences in pre-adolescents. *Foods* 9/9: 1307. https://doi.org/10.3390/foods9091307

Sick, Julia, Sara Spinelli, Caterina Dinnella and Erminio Monteleone. 2020b. Children's selection of emojis to express food-elicited emotions in varied eating contexts. *Food Quality and Preference* 85: 103953. https://doi.org/10.1016/j.foodqual.2020.103953

Siebenhaar, Beat. 2018. Funktionen von emoji und altersabhängigkeit ihres gebrauchs in der whatsapp-kommunikation. In Arne Ziegler ed. *Jugendsprachen: Aktuelle Perspektiven internationaler Forschung*. Berlin: De Gruyter, 749–772.

Siegel, Robert M., Amy Anneken, Christopher Duffy, Kenya Simmons, Michelle Hudgens, Mary Kate Lockhart and Jessica Shelly. 2015. Emoticon use increases plain milk and vegetable purchase in a school cafeteria without adversely affecting total milk purchase. *Clinical Therapeutics* 37/9: 1938–1943.

Solomon, Robert C. and Lori D. Stone. 2002. On "positive" and "negative" emotions. *Journal for the Theory of Social Behaviour* 32/4: 417–435.

Spina, Stefania. 2018. Role of emoticons as structural markers in Twitter interactions. *Discourse Processes* 56/4: 345–362.

Stark, Luke and Kate Crawford. 2015. The conservatism of emoji: Work, affect, and communication. *Social Media + Society* 1/2: 1–11.

Steinmetz, Katy. *2015. Oxford's 2015 Word of the Year is this Emoji*. *Time*. https://time.com/4114886/oxford-word-of-the-year-2015-emoji

Swaney-Stueve, Marianne, Tegan Jepsen and Grace Deubler. 2018. The emoji scale: A facial scale for the 21st century. *Food Quality and Preference* 68: 183–190.

Tang, Ying and Khe F. Hew. 2018. Emoticon, emoji, and sticker use in computer-mediated communications: Understanding its communicative function, impact, user behavior, and motive. In Liping Deng, Will W. K. Ma and Cheuk Wai Rose Fong eds. *New Media for Educational Change*. Singapore: Springer, 191–201.

Tang, Ying and Khe F. Hew. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International Journal of Communication* 13: 2457–2483.

Thurlow, Crispin, Christa Dürscheid and Federica Diémoz eds. 2020. *Visualizing Digital Discourse: Interactional, Institutional and Ideological Perspectives*. Berlin: De Gruyter.

Tigwell, Garreth W. and David R. Flatla. 2016. "Oh that's what you meant!": Reducing emoji misunderstanding. In Fabio Patternò, Kaisa Väänänen, Karen Church, Jonna Häkkilä, Antonio Krüger and Marcos Serrano eds. *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. Florence: ACM, 859–866.

Unicode. 2023. https://home.unicode.org

Upadhyay, Sri Siddhi N., Danielle N. Gunraj and Nicklas C. Phillips. 2023. Mad or mad-mad: Conveying subtle emotion with face emoji. *Frontiers in Psychology* 14: 1183299. https://doi.org/10.3389/fpsyg.2023.1183299

Verheijen, Lieke. 2016. Emoji voor dummies: Multimodaliteit in digitale communicatie met 144 pixels. *Over Taal* 55/3: 16–19.

Verheijen, Lieke. 2018. *Is Textese a Threat to Traditional Literacy? Dutch Youths' Language Use in Written Computer-Mediated Communication and Relations with their School Writing*. Nijmegen: Radboud University dissertation.

Wang, Yuan, Yukun Li, Xinning Gui, Yubo Kou and Fenglian Liu. 2019. Culturally-embedded visual literacy: A study of impression management via emoticon, emoji, sticker, and meme on social media in China. In Airi Lampinen, Darren Gergle and David A. Shamma eds. *Proceedings of the ACM on Human-Computer Interaction* 3: 1–24. https://doi.org/10.1145/3359170

Weissman, Benjamin. 2019. Peaches and eggplants or... something else? The role of context in emoji interpretations. *Proceedings of the Linguistic Society of America* 4/29: 1–6. https://doi.org/10.3765/plsa.v4i1.4533

Corresponding author
Lieke Verheijen
Radboud University
Department of Language and Communication
PO box 9103
NL-6500 HD
Nijmegen
The Netherlands
E-mail: lieke.verheijen@ru.nl

received: November 2023
accepted: January 2024

APPENDIX: EXAMPLE OF DATA COLLECTED

RiCL Research in Corpus Linguistics

Twitter conference discussion sessions: How and why researchers engage in online discussions

Rosana Villares
University of Zaragoza / Spain

Abstract – *Twitter* for academic purposes has been analysed from multiple perspectives such as genre analysis, the use of multimodality and hypertextuality, or type of participants; yet interactivity between writers and readers remains under-researched. This study analyses academic-related conversations from the *Twitter* conference genre, particularly focusing on the discussion session. Its objective is to identify the main interactional patterns, communicative functions, and digital discourse features in tweets. Dialogic turns were classified into comments, questions, responses, follow-up conversations, and automatic comments. Findings reveal that the main reasons behind online interaction correspond with community building and knowledge construction purposes. The digital medium does shape the form of tweets, which shows a high level of evaluative language, conversational style features, hedging, and emojis. All in all, these discursive features help create a welcoming and engaging style needed to engage in online science communication practices on social media.

Keywords – *Twitter*; discussion session; digital genres; communicative functions; digital discourse analysis; interactivity

1. INTRODUCTION[1]

During the last decade, there has been a growing need for researchers to adapt to new socioeconomic and cultural demands that reflect a shift in the creation, dissemination, and access to scientific knowledge. This is a consequence of Open Science policies advocating for a transparent and open sharing of research to expert and non-expert audiences (Luzón and Pérez-Llantada 2022). Some of the requirements that researchers face include gaining international visibility and recognition, meeting institutional standards, and securing public funding. To meet these evolving demands, researchers and

[1] This study was supported by the project *Digital Genres and Open Science (*PID2019-105655RB-I00 MCIN/AEI 10.13039/501100011033) funded by the *Spanish Ministry of Science and Innovation* and the *Spanish Agency for Research and the Government of Aragon* (H16_23). It is also a contribution to the *Erasmus + Project Digital Language and Communication Training for EU Scientists* (DILAN), co-funded by the *European Commission* (2022-1-ES01-KA220-HED-000086749). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may been made of the information contained therein.

scholars have embraced a range of digital genres that enable the dissemination of scientific knowledge to wide diversified audiences. Traditionally, 'genre' is understood as a communicative event with specific form conventions, targeting specific discourse communities, and fulfilling social actions (Miller 1984; Swales 1990). However, in the present digital landscape, researchers have access to digital resources and tools that enable them to share diverse data and findings with broader audiences, therefore relying on "new possibilities for interactivity and collaborative construction" of knowledge and participatory communication practices out of their discourse community (Belcher 2023: 38).

Relevant to digital genres in the context of Open Science is the notion of 'transformative science', which refers to the use of innovative online communication practices to disseminate scientific knowledge addressing the lay public and academic peers. Pérez-Llantada *et al.* (2022) reported that these transformative practices involve researchers using open-access repositories when sharing pre-prints and papers, academic social networks to stay updated on the latest developments in their respective fields, or social media platforms such as *Twitter* (*X*), *Facebook*, *Reddit*, and *Instagram* to communicate scientific research to special interest groups and broad audiences. It is well-known that among the various social media used by researchers, *Twitter* has gained prominence as a preferred medium for sharing scientific knowledge with both the lay public and scholarly peers due to its instant and short messaging nature. According to authors such as Büchi (2016), Lee *et al.* (2017), Côté and Darling (2018), Mehlenbacher (2019), and Tardy (2023), researchers join *Twitter* mainly to disseminate their work, promote their research outputs and publications, and network with colleagues in their disciplinary fields.

Research on *Twitter* usage by scholars has received attention because of its potential to make scientific knowledge accessible to diversified audiences (Darling *et al.* 2013; Lee *et al.* 2017; Luzón and Albero-Posac 2020). For instance, Darling *et al.* (2013) explored the usefulness of *Twitter* during the publication process as they analysed exchange practices among colleagues to generate ideas, receive peer-review comments, and increase the impact of their manuscripts' contents. Similarly, Lee *et al.* (2017) and Luzón and Albero-Posac (2020) investigated the practices of networking and communication in specific academic scenarios, particularly in academic conferences, where *Twitter* has become a powerful tool that combines with the on-site conferences as a means for

informal and formal communication. In this respect, Luzón and Albero-Posac (2020) identified four main communicative functions of conference tweets that have organisational purposes, promote informal interaction, foster community building, and focus on networking.

Further research on *Twitter* has focused on the rhetorical analysis of tweets written by scientists and public organisations (Orpin 2019; Tardy 2023), the use of digital affordances such as hyperlinking, multimodal composition, and intertextuality characteristic of academic tweets (Büchi 2016; Luzón 2023), or the different roles played by scientists when communicating science outside academia (Walter *et al.* 2019). Additionally, the analysis of the combination of different semiotic resources and *Twitter* affordances in tweet composition is a significant line of research in applied linguistics, where elements such as visuals, videos, hyperlinks, mentions, hashtags, and retweets, among others, are investigated to know how they help to disseminate messages and effectively engage wider audiences (Orpin 2019; Luzón and Albero-Posac 2020; Luzón 2023; Tardy 2023; Villares 2023a; Xu *et al.* 2023).

While previous research has tended to focus on multimodality and hypertextuality, the third key feature of digital genres —interactivity— has received less attention in the literature on *Twitter* for academic purposes. Interactivity on *Twitter* has examined the type of participants and readership of tweets (e.g., Walter *et al.* 2019), yet a deeper analysis from a discursive perspective has not been conducted yet. Tardy (2023) points out that scientific communication still occurs in its majority among audiences who are knowledgeable on the topics rather than reaching readership outside academia, so it seems relevant to examine how communicative exchanges between specialised audiences occur. To explore this issue, this paper analyses a corpus of academic tweets from an emerging digital genre called *Twitter* Conference Presentation (henceforth, TCP), which has remediated the traditional on-site academic conference presentation into the digital medium (Villares 2023a, 2023b). The TCP consists of a six-tweet thread where presenters share their research projects. Like on-site conference presentations, the TCP can be followed by a discussion session in the form of tweets that readers can post at the end of each thread to engage in a conversation with the presenter (and/or other readers). Based on the literature, it is hypothesised that even though the presenter cannot control who reads and responds to their content, they still want to initiate discussions and interact with potential readers. In order to give insights into how and why researchers may engage in

online conversations that take place on academic *Twitter*, the present study analyses the discussions following TCPs to address the following research questions:

1. Do *Twitter* Conference Discussion Sessions (henceforth TCDSs) follow the same interactional turn-type patterns as traditional on-site academic conference discussion sessions?
2. What are the main communicative functions and purposes of tweets in TCDSs?
3. Does the medium shape the type of digital discourse features participants use in their TCDS tweets?

Section 2 is devoted to an overview of the on-site conference discussion session. After that, Section 3 delves into the corpus description, data collection process, and analytical techniques. Section 4 reports the results in terms of turn-types, communicative functions supported by rhetorical strategies, and an exploration of digital discourse affordances. Finally, the paper concludes with a discussion of the main findings and their implications for researchers' communication skills development.

2. THE CONFERENCE DISCUSSION SESSION

The emergence and constant evolution of digital genres sometimes bring changes in the form, functions, and communicative purposes of traditional academic genres. In the case of the TCP, it still shares the primary communicative goals and functions of face-to-face conference presentations, that is, presenting work in progress and networking (Rowley-Jolivet and Carter-Thomas 2005; Hyland 2009), while introducing novel discursive and rhetorical strategies that arise from the affordances and constraints that *Twitter* (e.g., hashtags, mentions, retweets, or space restrictions) and the digital medium offer (Tagg 2015; Zappavigna 2017).

The conference presentation is part of a genre chain that consists of a series of genres organised in a chronological sequence. The conference presentation is preceded by genres such as the call for papers and the abstract, while it is followed by the conference paper and the discussion session (Räisänen 2002). The discussion session is defined as "the event that takes place right after a presentation at an academic conference in the form of dialogues between the 'presenter' and the 'discussants'" (Xu 2022a: 63). Regarding the organisation of interaction between participants in the discussion session,

Querol-Julián and Fortanet-Gómez (2012, 2014) identified its rhetorical and turn-taking structure, shedding light on its distinctive nature, and highlighting the combination of linguistic and non-linguistic features during the turn-taking exchanges. The three types of turns identified by the authors are: a) comments —when a turn includes a statement— b) questions —when a turn includes at least one question—, and c) responses from the presenters. In more detail, the generic structure of the dialogic exchange starts with the discussant's question followed by the presenter's turn (response). The question can include the following moves: a) announcing the question, b) asking the question, and c) reformulating the question. On the other hand, the presenter's response may consist of a reaction to the question, answering the question, expanding the topic of the question, and closing the turn (Querol-Julián and Fortanet-Gómez 2014: 86).

Concerning the main communicative functions of turns, discussant turns are characterised by an evaluative function and specific lexico-grammatical features attached to that function. Drawing upon Webber's (2002) comprehensive account of question types and participants' reactions during discussion sessions, it is possible to classify question functions into five main categories: a) fact-seeking questions, b) opinion-seeking questions, c) justification-seeking questions, d) suggestions, and e) neutral statements. There is a gradual evaluative function in the different turns, ranging from low evaluative turns (e.g., facts or statements) to high evaluation when criticism appears (e.g., justification-seeking questions). In order to reduce the potential threat of criticism in a turn, Xu (2022a) argued that building rapport was used as a common practice among discussants who dedicated more effort to thank and praise the presenter at the beginning of their turn. Rhetorical strategies to soften criticism can take the form of hedges (e.g., *I think*), admission of limitations (e.g., *I don't know*), and evaluative language to show appreciation (Webber 2002; Hyland 2005; Xu 2022b). Hence, rapport-building strategies that contextualise and introduce a comment rely on discursive features of politeness and solidarity, alongside other lexico-grammatical features, which tend to be employed when there is a high level of evaluation (Wulff *et al.* 2009; Xu 2022a). Additionally, Konzett (2012) —in her book about identity construction at academic conferences— noted that the purpose of raising a question may extend beyond seeking or evaluating scientific information, to encompass aspects of negotiating professional identities and self-promotion. In sum, a discussant's turn may include different communicative purposes other than seeking and exchanging knowledge.

3. METHODS AND PROCEDURE

3.1. Corpus description

The corpus comprises 561 tweets (13,105 words) posted in the discussion sessions of 55 presentations in English[2] from the *Twitter Conference Linguistweets* (ABRALIN 2020).[3] Yet, the corpus is multilingual and includes tweets in English (81%), French (10%) Portuguese (5%), and other languages (4%). There is a total of 235 discussions. On average, each presentation received ten tweets and up to three users interacted in the discussion. Examining the composition of the corpus in more detail, as shown in Table 1, tweets were organised into 'comments', 'questions', 'responses', 'automatic comments', and 'follow-up conversations' whenever a discussion included more than two tweets.

Turn-type		**Number of tweets**	**Percent**
Automatic comment		2	0%
Comment		158	28%
Question		89	16%
Response		169	30%
Follow-up conversation		143	25%
	Follow-up comment	77	14%
	Follow-up question	19	3%
	Follow-up response	47	8%
Total		**561**	**100%**

Table 1: Corpus description

Questions and comments are tweets written by readers that initiate the discussion, while responses are the presenters' replies. Sometimes a response can consist of more than one tweet. The category follow-up conversation refers to a discussion that involves more than just the standard turn-taking sequence of comment-response or question-response. 25 per cent of the corpus tweets belonged to this category, which involved longer interactions, where participants engaged in longer exchanges repeating the comment-response pattern. Finally, a medium-related category was identified —'automatic comments'— which consisted of tweets automatically generated by a software to promote presentations.

[2] A list with the presentation titles and links can be found in Appendix 1.
[3] https://abralin.org/es/evento/linguistweets-3/

3.2. Corpus collection

Tweets were accessed at the conference website (https://www.linguistweets.org/linguistweets-2020/programa/), manually downloaded from the participants' *Twitter* accounts, and stored in *Word* documents so that both text and other semiotic resources (e.g., images, emojis) could be analysed. 55 documents were compiled to store the TCDSs separately and filed under an anonymous name, e.g., TCDS1 for the tweets comprising all the discussions associated with the first presentation. For each TCDS, the following items were identified: title of the presentation and hyperlink, number of turns, and participants (presenter and readers). If different readers commented on one presentation, readers were labelled as Reader 1, Reader 2, Reader 3, and so on, to track the different discussions that took place.

3.3. Analytical procedure

The TCDS documents were uploaded to the qualitative data software *ATLAS.ti*. version 8.4.5.[4] The codification of the corpus started with an inductive approach based on a close reading of the corpus tweets. To assure the reliability of the annotation process, the author used the memo and code description options of *ATLAS.ti*, which allow describing in detail coding procedures and decisions that help to guarantee consistency during the labelling process (Krippendorff 2004; Paulus 2022). The coding system was revised and redefined in three cycles to reach a saturation point of codes and carried out in three-time intervals to guarantee the validity of the codification (Saldaña 2009).

The coding cycle began with the identification and description of the technical and discursive features shaped by the medium (e.g., *Twitter* affordances and constraints, digital discourse features). In particular, I focused on multimodal semiotic resources (Orpin 2019; Luzón and Albero-Posac 2020; Luzón 2023), *Twitter* formal elements (Luzón 2023; Tardy 2023), and linguistic features common in digital discourse (Mauranen 2013; Tagg 2015; Zappavigna 2017; Luzón and Albero-Posac 2020). Regarding the latter, previous studies on face-to-face discussion sessions have analysed non-linguistic resources such as gestures, facial expressions, loudness, or laughter (Wulff *et al.* 2009; Querol-Julián and Fortanet-Gómez 2012). Transferring these features into the digital medium can be done through visual resources (emojis, smileys), punctuation

[4] https://atlasti.com/

(exclamation marks, capitalisation), or characteristic features of spoken dialogue (addressing interlocutors by their first names, interjections, laughter, or lengthened vowels). Lastly, I examined features of interpersonality as established in Hyland's (2005) stance and engagement framework, to pinpoint what strategies were used by presenters and readers to interact with one another (Querol-Julián and Fortanet-Gómez 2012; Orpin 2019; Luzón and Albero-Posac 2020; Luzón 2023).

The communicative functions of tweets were also coded. For readers' tweets, i.e., questions and comments, I followed Xu's (2022a) taxonomy of questions (fact-seeking, opinion-seeking, justification-seeking, suggestion-making, and comment) for an initial overview of communicative functions. Regarding questions, 95 percent were contextualised questions, which meant that in addition to the question itself, other moves such as announcing the question, greeting, praising, or sharing some personal information relevant to the situation before posing the question were needed, a similar situation to what others had previously noted in their analysis of face-to-face discussion sessions (Querol-Julián and Fortanet-Gómez 2014; Xu 2022a). After a reiterative process of re-reading tweets and a redefinition of codes, 32 communicative functions were identified and classified into three main communicative purposes: knowledge construction, community building, and self-promotion. Table 2 summarises the categories and codes of the coding system.

Category	**Codes**
Multimodal resources	Emojis, gifs, images, smileys, videos.
Twitter formal elements	Embedded tweets, hashtags, hyperlinks, mentions.
Digital discourse features	Abbreviations, capitalisation, contractions, intensifiers (adverbs, repetition of words/symbols), exclamation marks, interjections, laughter, lengthened vowels.
Interpersonality	Stance: Self-mentions (first person pronouns, possessives), hedges (modal verbs and conversational hedges e.g., *just*, *a little bit*), attitude markers (evaluative adjectives, verbs).
	Engagement: Reader mentions (second person pronouns, possessives, vocatives), personal asides.

Table 2: Description of the coding system

Category	Codes
Knowledge construction	Acknowledging collaboration, acknowledging limitations, agreeing with a previous idea, asking for feedback, discussing an idea, exemplifying, explaining content, making requests, making suggestions, offering a neutral statement, referring to previous studies, requesting clarification, requesting an opinion, seeking factual information, seeking justification, sharing resources.
Community building	Addressing the reader, apologising, appraising the presenter's work, conveying gratitude, down-toning, engaging in humour, expressing politeness, expressing strong feelings, greeting, keeping in touch after the conference, sharing personal information, sharing research interests.
Self-promotion	Expressing significance, promoting one's outputs, raising awareness, referring to future work.

Table 2: (Continuation)

4. RESULTS

The results show that the most frequent communicative purposes of TCDS tweets are community building and knowledge construction. These findings are reported in Section 4.1., where communicative functions and rhetorical strategies are analysed in relation to the different turn-types of TCDSs. Section 4.2. reports on the digital medium-related characteristics of TCDSs.

4.1. Communicative functions and discursive realisations of turns

4.1.1. Comments

Comments often take the form of statements written by the reader and are related to an interpersonal or community-building dimension, hence, granting more importance to building rapport and interpersonal relations than to knowledge construction (Table 3).

Community building (N=208)	
Communicative Function	**Frequency**
Appraising the presenter's work	83
Addressing the reader	22
Conveying gratitude	21
Engaging in humour	20
Expressing strong feelings	20
Sharing personal information	15
Sharing research interests	13
Keeping in touch after the conference	7
Apologising (for a mistake)	3
Greeting	2
Down-toning	1
Expressing politeness	1

Table 3: Frequencies of communicative functions in comments.

Knowledge construction (N=99)	
Communicative Function	**Frequency**
Offering a neutral statement	40
Making suggestions	15
Referring previous studies	9
Explaining content	8
Exemplifying	7
Sharing resources/outputs	6
Agreeing with a previous idea	4
Seeking factual information	4
Acknowledging collaboration	2
Making requests	2
Acknowledging limitations	1
Asking for feedback	1
Self-promotion (N=26)	
Communicative Function	**Frequency**
Expressing significance	11
Promoting one's outputs	9
Raising awareness	6

Table 3: Continuation

Comments conveyed the following functions: a) appraising the presenter's work, b) offering a neutral statement, c) addressing the reader, d) conveying gratitude, e) engaging in humour, and f) expressing strong feelings. All of them work to establish rapport and a positive evaluation of the conversation that takes place. Examples (1–6)[5] illustrate the different functions:

(1) Super interesting presentation! Thank you! I've never thought about memes as giving advice before, but it makes sense. I'll be keeping my eyes out for that now. (TCDS6_Reader3)

(2) I've heard my dad & uncle (both from Michigan, USA but raised by two Appalachian English–speaking parents) say things like "They wanted to get married real quick" to mean "they wanted to get married in a short amount of time." For me though I can't do this with postverbal "quick" (TCDS14_Reader1)

(3) Really enjoyed this Martin! So clear and fun :) always a pleasure to read your stuff, greetings from NZ (TCDS26_Reader5)

(4) Thanks for the paper! and the refs! (TCDS23_Reader1)

(5) Meow Viry Much Madame ! We potitchats from France are realy proud /20 (TCDS45_Reader1)

(6) Awwww we lost! we (renov) were winning at one point. (TCDS26_Reader1)

Comments correlate with appraising the presenter's work at the beginning of the tweet (1–4) and sharing personal information and research interests (1–2). At the textual level,

[5] Examples are verbatim transcriptions of tweets. Tweets not written in English include the translation in brackets.

a high frequency of self-mentions through the first-person pronouns *I* (1–2) or *we* representing a group is found (5–6). Subject omission is also observed (3), as part of a more conversational register, which is also noticed in other linguistic digital discursive features such as the use of exclamation marks, especially after thanking or requesting, abbreviations and contractions, spelling mistakes, letter repetition, interjections, or the use of vocatives.

Evaluative language, in particular positive evaluation, is a common trait in TCDSs through the use of intensifiers (*super*, *really*) and adjectives describing the presentation's contents, as shown in (1), (3), and (5). The use of these strategies creates a close bond between the presenter and the reader, both relying on politeness strategies to create rapport by positively appraising the presenter, addressing the reader directly, and conveying gratitude (3–4). These communicative functions reflect a focus on the person rather than on the (scientific) content. Similarly, the use of humour in (5) is another strategy that can include inside-group jokes and references to shared interests between the presenter and the reader. As found by Wulff *et al.* (2009) in their analysis of laughter in conference discussion sessions, laughter and humour tend to be present to soften potential criticism or requests, or to break the ice at the beginning of a conversation.

4.1.2. Questions

Table 4 shows the distribution of communicative purposes with their corresponding communicative functions in questions. The main purpose of questions is to construct and exchange knowledge, closely followed by community building.

Community building (N=121)	
Communicative Function	**Frequency**
Appraising the presenter's work	40
Conveying gratitude	15
Expressing politeness	15
Addressing the reader	11
Expressing strong feelings	6
Greeting	5
Sharing research interests	5
Engaging in humour	3
Down-toning	1

Table 4: Frequencies of communicative functions in questions

Knowledge construction (N=127)	
Communicative Function	**Frequency**
Seeking factual information	48
Requesting an opinion	25
Making requests	13
Making suggestions	12
Requesting clarification	8
Referring previous studies	6
Exemplifying	5
Sharing resources/outputs	5
Seeking justification	4
Agreeing with a previous idea	1
Self-promotion (N=1)	
Communicative Function	**Frequency**
Referring future work	1

Table 4: Continuation

As far as questions are concerned, the main communicative functions they fulfil are four: a) seeking factual information, b) appraising the presenter's work, c) seeking opinion, and d) expressing politeness. Two out of the three most frequent functions coincide with Xu's (2022a) taxonomy of conference questions as illustrated in:

(7) Are there books that use the same font for both? (TCDS3_Reader1)

(8) Beautiful graph! What did you use to do that? (TCDS5_Reader1)

(9) Thank you for this talk! We agree, I think. We looked into German and Dutch a little and wondered about the distinction between the subordinating and coordinating becauses. Do you have any thoughts on that? (TCDS55_Reader1)

Example (7) illustrates a straight question that is purely fact-seeking, but examples (8) and (9) show how readers prefer to contextualise the question before making a request. For instance, questions are introduced first by praising the presentation or a specific part of the presentation with positive evaluative adjectives (8) or by congratulating and sharing some research interests that position the reader at the same level as the presenter in terms of knowledge (9). Likewise, when requesting an opinion from the presenter, in addition to addressing the presenter directly with the pronoun *you*, readers often appraise their work and convey gratitude using the same linguistic resources and standard formulaic politeness strategies (full grammatical sentences, polite requests, hedging) before posing a question that could be interpreted as threatening to the presenter's expertise.

4.1.3. Responses

The main communicative purposes of *response* tweets by the presenter are knowledge construction and community building. Regarding the third communicative purpose, self-promotion, it has the highest occurrence in the response category (Table 5).

Community building (N=158)	
Communicative Function	**Frequency**
Conveying gratitude	70
Addressing the reader	19
Appraising the presenter's work	14
Keeping in touch after the conference	14
Engaging in humour	9
Expressing strong feelings	8
Sharing personal information	6
Greeting	5
Apologising (for a mistake)	4
Sharing research interests	4
Down-toning	3
Expressing politeness	2
Knowledge construction (N=197)	
Communicative Function	**Frequency**
Explaining content	83
Acknowledging limitations	25
Agreeing with a previous idea	25
Exemplifying	20
Referring previous studies	14
Sharing resources/outputs	14
Requesting clarification	6
Making requests	5
Acknowledging collaboration	2
Asking for feedback	2
Requesting an opinion	1
Self-promotion (N=30)	
Communicative Function	**Frequency**
Promoting one's outputs	16
Referring future work	13
Expressing significance	1

Table 5: Frequencies of communicative functions in responses

Presenter response tweets realise five communicative functions in the data: a) explaining content, b) conveying gratitude, c) agreeing with previous ideas, d) acknowledging limitations, and e) exemplifying, as shown in:

(10) Yes, we looked at all loanwords that occurred at least five times across our corpus, regardless of their meaning, and we do find loanwords within the same text that are not semantically related, e.g. paua (shell) and aroha (love) :) (TCDS5_Presenter)

(11) Thank you! Yes, I think they all derive from the original POSS.2SG along the path: possessive > salient > anaphoric > proprial Although, I'm not exactly sure, how the last link works. I hope to get a chance to present these hypothesis at #SLE2021 (TCDS21_ Presenter)

(12) I am definitely not! I fully agree with you and thank you for pointing this out. This was just to make it easy for people to know what I was talking about in the first tweet. We were limited to six! 😳 (TCDS45_ Presenter)

Aligning with the nature of responses in face-to-face discussion sessions, their main communicative function in TCDSs is explaining by elaborating on content to answer questions (10) and (11). Explanations often appeared in combination with exemplification, a strategy used by presenters to illustrate abstract concepts. Similar to questions and comments, conveying gratitude by thanking the other person (for either reading the presentation or posing a question) could be considered an obligatory function in view of its frequent use in responses. As part of a friendly and polite environment, many responses agreed with previous comments by readers (10–12). These functions are realised with exclamation marks after *thanks* or *thank you* to stress friendliness and enthusiasm, emojis and smileys, evaluative language (*I fully agree*), and the use of first person-pronouns that make the presenter's voice visible (10–12).

Another significant function is the authors' acknowledgment of the limitations of their research by hedging (11) or down-toning, as in (12), where the presenter acknowledges that the presentation content has been simplified because of space constraints. This positioning shows presenters not as knowledge holders but rather as participants in the knowledge construction process.

Lastly, the self-promotion communicative purpose, even though it occurs less frequently than the community building and knowledge construction purposes, occurs most frequently in responses, especially through the function of promoting one's publications and outputs. This might result from the fact that presenters are expected to provide references for their presentations' contents.

4.1.4. Follow-up conversations

When the conversation between reader and presenter continued after the presenter's response, community building and knowledge construction continued to be the most relevant purposes of longer interactions. As shown in Table 6, the main communicative functions of tweets were: a) conveying gratitude, b) agreeing with a previous idea, c) explaining content, d) appraising the presenter's work, e) expressing strong feelings, and f) keeping in touch after the conference.

Community building (N=161)	
Communicative Function	**Frequency**
Conveying gratitude	43
Appraising the presenter's work	29
Expressing strong feelings	18
Keeping in touch after the conference	17
Engaging in humour	16
Sharing personal information	13
Down-toning	7
Expressing politeness	6
Sharing research interests	5
Apologising (for a mistake)	4
Addressing the reader	3
Knowledge construction (N=155)	
Communicative Function	**Frequency**
Agreeing with a previous idea	35
Explaining content	31
Referring previous studies	13
Acknowledging limitations	12
Discussing an idea	12
Sharing resources/outputs	11
Exemplifying	10
Making requests	9
Seeking factual information	8
Making suggestions	6
Offering a neutral statement	3
Requesting an opinion	2
Requesting clarification	2
Asking for feedback	1
Self-promotion (N=13)	
Communicative Function	**Frequency**
Promoting one's outputs	7
Referring future work	5
Expressing significance	1

Table 6: Frequencies of communicative functions in follow-up conversation

As follow-up conversations consist of several tweets between readers and presenters, in particular further comments and responses, communicative functions such as conveying gratitude and positively appraising the presenter's work are commonly found. With this turn-type, the conversation topic is expanded, so agreeing with previous ideas presented in the tweets is a frequent function. Agreeing is also used as a positive politeness strategy to foster bonds between readers and presenters. However, whenever the reader insists on the topic, the use of hedges (*I was just curious*) and other polite strategies (*I was wondering if you could have*, *but maybe this is already moving too far*) as well as relying on specific resources (*I'll quickly look at my data*, *I'm going to look up articles*) are frequent to justify their questions and answers, as illustrated in (13) and (14):

(13) Reader2: Great thread!! I was just curious about what you make of l.14? 😄

Presenter: Thank you! :) I consider l. 14 has a confirmation and acceptance of the proposed other-increments. I find it interesting how Anna uses overt dependent syntax ("che") so her talk is dependent on Paolo's... but Paolo's turn l. 14 is an independent clause. ;)

Reader2: Thank you! Yeah, I was wondering if you could have cases in which the acceptance is also designed as dependent sort of recycling the increment. But maybe this is already moving too far from your point here hehe. Anyway, great job!

Presenter: No, but it's a great question. It allows me to think about the notion of recycling/repeating in relation to acceptance/confirmation of a candidate! I'll quickly look at my data and come back to you! :D thank you so so much! (TCDS39_Follow-upConversation)

These polite strategies are interwoven with digital discourse features like exclamation marks, laughter, repetitions, and contractions to seem friendlier. In a similar vein, emojis and smileys are more frequent in this turn once that contact has already been established between users.

(14) Reader2: This is super interesting, thank you! In my research I look at the link between acquisition and language change. Would you say this could link up with a cue-based approach to change [...] ½

Reader2: i.e children are sensitive to prosodic cues in their input and this can cause frequency changes for linguistic forms, leading to overall language change?

Presenter: I don't know much about the link between language acquisition and change, though I find it a very interesting subject. But I would guess yes, since prosodic cues are so important for language acquisition, they could also influence language change through changes in input cues.

Reader2: I wonder if there's any evidence out there that's shown change occurring in synchronic acquisition data. I know Marit Westergaard works on syntactic cues a lot, so was just wondering if it extended to prosody! Thanks!

Presenter: Thank you for the question! I wish I could help you more, but unfortunately I can't think of any study on this subject. If I do think of something, I'll let you know :)

Reader2: No problem at all! Thanks for sharing your research. Likewise, I think I'm now going to look up articles related to prosodic cues and change so will let you know 😃 (TCDS34_follow-upConverstation)

Likewise, expressing strong feelings through evaluative language (e.g., *great*, *cool*, *I'd love to*), and using adverbial intensifiers (*super*, *so*), reader mentions (*you*, *your interest*) or conditional sentences (*if I do think of…*, *I wish I could…*) might be associated with various communicative functions such as keeping in touch after the conference,

exchanging resources, or discussing ideas. All in all, participants try to come across as friendly and supportive.

4.2. Digital discourse features of TCDSs

In view of TCDSs taking place on a digital platform, communicative practices from physical discussion sessions can be digitally remediated or new practices might emerge resulting from *Twitter*'s affordances and constraints. Table 7 shows the frequency and distribution of *Twitter* formal elements (embedded tweet, hashtag, mention), hyperlinking, and multimodal assemblage of semiotic resources (image, gif, emoji/smileys). Total frequencies are broken down by turn-type. The features that stand out the most in TCDSs are emojis/smileys, mentions, and hyperlinks.

Features	**Comment**	**Question**	**Response**	**Follow-up**	**Total**
Embedded tweet	1	1	2	1	5
Hashtag	12	3	2	1	18
Mention	24	8	5	1	38
Hyperlink	2	4	9	5	20
Total *Twitter* formal elements	**39**	**16**	**18**	**8**	**81**
Emoji/smiley	47	11	58	47	163
Gif	0	0	0	1	1
Image	1	0	0	1	2
Total multimodal elements	**48**	**11**	**58**	**49**	**166**

Table 7: Distribution of *Twitter* formal elements and multimodal elements by turn-type

In opposition to TCP, which are heavily loaded with images (Villares 2023a), visual elements such as images and gifs are scarce in TCDSs. Only emojis and smileys are used frequently by both readers and presenters. As identified in previous sections, the main function of emojis is to show the attitude of the reader or presenter, which corresponds with either a positive evaluation that helps to create a sense of closeness and friendliness, as in (10) and (12–14), or to express concerns when acknowledging limitations or explaining content, (15–16):

(15) Possibly, but these stigmatized variants are mostly associated with rural areas. Perhaps they have been lost in those rural areas, or perhaps they are not as stigmatized anymore because of migration. That's something we wonder now 🤔🤷‍♀️ ... It's too small a study to know for sure at this point! (TCDS4_Presenter_Follow-upConversation)

(16) I don't know if reviewer #2 will accept our conclusions on the basis of a twitter poll, but I'm sure glad it's in line with our theory. 😅 Across languages, the suprasegmental rules of clipping vary substantially. (TCDS26_Presenter_Follow-upConversation)

Regarding the praising and conveying gratitude functions, sometimes, tweets with no text, only emojis (e.g., 👏👏👏, 👌🏻, 😻), could be found. They frequently represented a paralinguistic feature common in the discussion session, i.e., clapping, agreeing, or smiling. In other cases, emojis appear alongside text, reinforcing or enhancing the tweet's communicative function.

Moving on to *Twitter*'s specific formal elements, mentions were the preferred resource. Like vocatives, mentions tend to appear at the beginning of the tweet, often creating a sense of proximity, and allowing immediate interaction because it notifies and explicitly addresses other users. Mentions are followed by communicative functions such as expressing strong feelings, expressing gratitude, or praising the presenter's work as illustrated in (17):

(17) Trabalho maravilhoso, @ NicolaDaly18! A multimodalidade tem se mostrado uma excelente aliada no ensino-aprendizagem de línguas. (Wonderful study, @NicolaDaly18! Multimodality has proved to be an excellent ally in the learning and teaching of languages) (TCDS3_Reader2_comment)

Other uses of mentions can refer to a presenter naming co-authors (18), calling out the author of a resource that could be useful within the discussion (19), or informing a third person of the existence of the presentation (20):

(18) Co-author on this work is @elles_belles (who I didn't tag because she never tweets haha) (TCDS31_Presenter_comment)

(19) @uhlon dohlenko wrote a paper about anglophone lolspeak a loooong time ago! (I want to say "before it was cool", but I guess that was actually at the height of its popularity?) (TCDS45_Reader3_comment)

(20) @pbcardoso, see this! (TCDS36_Reader3_comment)

Hyperlinks are another digital affordance that fulfils the communicative function of sharing resources and promoting one's outputs or publications (e.g., links to a paper's DOI and repositories) and sharing resources (e.g., software, code, websites) by both presenters and readers. The remaining *Twitter* technical elements are hashtags, which are mainly used to relate the content of tweets to the conference (e.g., *#SLE2021*) or for humoristic purposes (e.g., *#SauronEye*, *#SavetheGricean*). Embedded tweets work as referencing tools so that participants can point to specific information mentioned during presentations to share outputs/resources or as promotional tools that increase the presentation's visibility when it is shared in other discussions.

5. CONCLUSION

This study has explored how TCDSs remediate the face-to-face academic conference discussion session digitally. Both genres have a similar sequential organisation beginning with a comment or question posed by a reader and finishing with the presenter's response (Querol-Julián and Fortanet-Gómez 2014). However, the digital genre introduces some novelty, particularly when interaction exceeds two turns (i.e., tweets). This turn-type, which I labelled 'follow-up conversation', consists of both readers and presenters elaborating on their answers and expanding the conversation to areas of knowledge construction, community building, and self-promotion. The automatic comment, which refers to promotional software-generated tweets, was also considered a medium-related turn. Regarding readers' turns, while TCDS questions can be either straightforward or contextualised (Xu 2022a), most tweets were contextualised, which fostered a bond between participants by drawing on different community-building communicative functions.

Regarding the communicative purposes of TCDSs, participants engage in conversation to establish interpersonal relationships (community building), exchange knowledge (knowledge construction), and to a lesser extent, self-promotion. Hence, TCDS participants engage in conversations with similar communicative purposes as users of other digital genres such as science blogs, academic conference tweets, or tweetorials (Mauranen 2013; Luzón and Albero-Posac 2020; Tardy 2023). Community-building communicative functions are means to establish interpersonal relationships between presenters and readers in a positive and polite manner. Compared to face-to-face conferences, paralinguistic strategies (e.g., gestures, body language, facial expressions) are remediated in the digital medium with the adoption of informal digital discourse. By using vocatives, exclamation marks, emojis, or evaluative language, participants create rapport and a friendly environment. Knowledge-construction communicative functions commonly include an exchange of specific information and opinions that should be explained and justified with examples, data, or references. This often occurs in follow-up conversations because they grant more space to delve into the topics, while face-to-face discussions tend to give hush answers due to time constraints. Moreover, *Twitter* allows participants to use its affordances, i.e., hyperlinking and multimodality, to give richer answers and move discussions forward.

In opposition to what previous academic *Twitter* research suggests (Büchi 2016; Lee *et al.* 2017; Côté and Darling 2018; or Mehlenbacher 2019), self-promotion is not a relevant communicative purpose for participants during the discussion. This finding contrasts with the TCP, the previous genre to the TCDS. In TCPs, presenters focus on knowledge construction and self-promotion reflected on the use of discursive strategies such as questions, informative images, or semantic hashtags to signal key terms to make the presentations attractive and informative (Villares 2023a, 2023b). In the case of TCDSs, however, there is a focus on community-building and networking, therefore, relying on a conversation style characterised by emojis/smileys, exclamative sentences, and evaluative language. These discursive features are enhanced by the digital medium and imitate both linguistic and paralinguistic features of face-to-face interaction.

This study presents some limitations that should be commented on. The study's data come from a small corpus that should be expanded to include a larger corpus of academic tweets, either from TCDS or other *Twitter*-related genres such as publication-promoting tweetorials or other academic-related threads (Luzón 2023; Tardy 2023). Likewise, the analytical framework could be applied to other digital genres that promote interaction among diversified audiences such as *Reddit* forums or citizen science websites to test its efficacy in analysing science communication practices. A second limitation refers to the fact that the data analysis was carried out by just one person. Even though contingency measures were implemented to ensure consistency and the validity of results with *ATLAS.ti* tools, it is advisable for future studies to involve more researchers who could ensure high inter-rater reliability agreement levels. Thirdly, from a methodological perspective, the discursive analysis of tweets could have been complemented with interviews or questionnaires to some of the presenters and readers to validate the study's results. This action would have shed light on the participants' actions and intentions, and it might open a future avenue for research.

Finally, as this paper has described digital communicative practices of international academic communities, the findings may have some pedagogical implications. Nowadays researchers find themselves in a paradigm where science needs to be communicated in a transparent, accessible, and engaging way, yet few opportunities to learn and develop this skill are offered by their institutions. To achieve this goal, it is crucial to understand how digital science communication happens so that research-based training shows scholars how to share scientific knowledge, prompt discussions, or foster collaboration among

diversified audiences in the new digital genres. Hence, this study contributes to the current research on the identification of new communicative practices within the framework of digital genre analysis and the importance of social media for community building and knowledge construction.

REFERENCES

Belcher, Diane D. 2023. Digital genres: What they are, what they do, and why we need to better understand them. *English for Specific Purposes* 70: 33–43.

Büchi, Moritz. 2016. Microblogging as an extension of science reporting. *Public Understanding of Science* 26/8: 953–968.

Côté, Isabelle M. and Emily S. Darling. 2018. Scientists on Twitter: Preaching to the choir or singing from the rooftops? *FACETS* 3/1: 682–694.

Darling, Emily S., David Shiffman, Isabelle M. Côté and Joshua A. Drew. 2013. The role of Twitter in the life cycle of a scientific publication. *Ideas in Ecology and Evolution* 6: 32–43.

Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2: 173–192.

Hyland, Ken. 2009. *Academic Discourse: English in a Global Context*. New York: Continuum.

Konzett, Carmen. 2012. *Any Questions? Identity Construction in Academic Conference Discussions*. Berlin, Boston: De Gruyter Mouton.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to its Methodology*. Thousand Oaks: Sage Publications.

Lee, Mi Kyung, Ho Young Yoon, Marc Smith, Hye Jin Park and Han Woo Park. 2017. Mapping a Twitter scholarly communication network: A case of the association of internet researchers' conference. *Scientometrics* 112/2: 767–797.

Luzón, María José. 2023. Multimodal practices of research groups in Twitter: An analysis of stance and engagement. *English for Specific Purposes* 70: 17–32.

Luzón, María José and Sofía Albero-Posac. 2020. 'Had a lovely week at #conference2018': An analysis of interaction through conference tweets. *RELC Journal* 51/1: 33–51.

Luzón, María José and Carmen Pérez-Llantada. 2022. *Digital Genres in Academic Knowledge Production and Communication: Perspectives and Practices.* Bristol: Multilingual Matters.

Mauranen, Anna. 2013. Hybridism, edutainment, and doubt: Science blogging finding its feet. *Nordic Journal of English Studies* 12/1: 7–36.

Mehlenbacher, Ashley Rose. 2019. *Science Communication Online: Engaging Experts and Publics on the Internet*. Columbus: The Ohio State University Press.

Miller, Carolyn R. 1984. Genre as social action. *Quarterly Journal of Speech* 70: 151–167.

Orpin, Deborah. 2019. #Vaccineswork: Recontextualizing the content of epidemiological reports on Twitter. In María José Luzón and Carmen Pérez- Llantada eds. *Science Communication on the Internet: Old Genres Meet New Genres*. Amsterdam: John Benjamins, 173–194.

Paulus, Trena. 2022. Digital tools for digital discourse analysis. In Camilla Vásquez ed. *Research Methods for Digital Discourse Analysis.* New York*:* Bloomsbury Publishing, 115–137.

Pérez-Llantada, Carmen, Olga Abián, Cristina Cadenas-Sánchez, Oana Carciu, Jesús Clemente-Gallardo, Idoia Labayen, Bienvenido León, Maria Carmen Erviti, Alfonso Ollero, Maddi Oses Recalde, Diego Rivera, Alberto Vela, Adrian Velazquez-Campoy, Rosana Villares and Ana Cristina Vivas Peraza. 2022. *Digital Science: Sustainable, Transformative and Transversal. Final Report*. Mendeley Data V1. https://doi.org/10.17632/2yv5brwxg5.1

Querol-Julián, Mercedes and Inmaculada Fortanet-Gómez. 2012. Multimodal evaluation in academic discussion sessions: How do presenters act and react? *English for Specific Purposes Journal* 3/4: 271–283.

Querol-Julián, Mercedes and Inmaculada Fortanet-Gómez. 2014. Evaluation in discussion sessions of conference presentations: Theoretical foundations for a multimodal analysis. *Kalbotyra* 66: 77–98.

Räisänen, Christine. 2002. The conference forum: A system of interrelated genres and discursive practices. In Eija Ventola, Celia Shalom and Susan Thompson eds. *The Language of Conferencing*. Berlin: Peter Lang, 69–93.

Rowley-Jolivet, Elizabeth and Shirley Carter-Thomas. 2005. The rhetoric of conference presentation introductions: context, argument and interaction. *International Journal of Applied Linguistics* 15/1: 45–70.

Saldaña, Johnny. 2009. *The Coding Manual for Qualitative Researchers*. Thousand Oaks: Sage Publications.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Tagg, Carolyne. 2015. *Exploring Digital Communication: Language in Action*. New York: Routledge.

Tardy, Christine M. 2023. How epidemiologists exploit the emerging genres of Twitter for public engagement. *English for Specific Purposes* 70: 4–16.

Villares, Rosana. 2023a. Twitter conference presentations: A rhetorical and semiotic analysis of an emerging digital genre. *ELIA: Estudios de Lingüística Inglesa Aplicada* 22: 125–167.

Villares, Rosana. 2023b. Exploring rhetorical strategies of stance and engagement in Twitter conference presentations. *ESP Today* 11/2: 280–301.

Walter, Stephanie, Ines Lörcher and Michael Brüggemann. 2019. Scientific networks on Twitter: Analyzing scientists' interactions in the climate change debate. *Public Understanding of Science* 28/6: 696–712.

Webber, Patrick. 2002. The paper is now open for discussion. In Eija Ventola, Celia Shalom and Susan Thompson eds. *The Language of Conferencing*. Berlin: Peter Lang, 227–253.

Wulff, Stefanie, John M. Swales and Kristen Keller. 2009. "We have seven minutes for questions": The discussion sessions from a specialized conference. *English for Specific Purposes* 28: 79–92.

Xu, Xiaoyu. 2022a. A genre-based analysis of questions and comments in Q&A sessions after conference paper presentations in computer science. *English for Specific Purposes* 66*:* 63–76.

Xu, Xiaoyu. 2022b. Differences between novice and experienced academics in their engagement with audience members in conference Q&A sessions. *Journal of English for Academic Purposes* 60: 101188. https://doi.org/10.1016/j.jeap.2022.101188

Xu, Xiaoyu, Jerois Gevers and Luca Rossi. 2023. “Can I write this is ableist AF in a peer review?”: A corpus-driven analysis of Twitter engagement strategies across disciplinary groups. *Ibérica* 46: 207–236.
Zappavigna, Michelle. 2017. Twitter. In Christian Hoffmann and Wolfram Bublitz eds. *Pragmatics of Social Media*. Berlin: De Gruyter Mouton, 201–224.

Corresponding author
Rosana Villares
University of Zaragoza
Department of English and German
Facultad de Economía y Empresa
Gran Vía, 6
Zaragoza 50005
Spain
E-mail: rvillares@unizar.es

received: November 2023
accepted: July 2024

APPENDIX 1

1. Citing linguistic data: The Tromsø Recommendations:
   https://twitter.com/superlinguo/status/1335011220152229888
2. What's in a name?
   https://twitter.com/sheeli3/status/1335015044745191431
3. The Linguistic Landscape of Bilingual Picturebooks:
   https://twitter.com/NicolaDaly18/status/1335023291937947648
4. Value judgments associated with allophones of alveolar tap and trill:
   https://twitter.com/porraschaver/status/1335029874285633540
5. Exploring Loanword Networks:
   https://twitter.com/TryeDavid/status/1335037615326674944
6. Tempering and Aligning Advice-Giving Through Memes:
   https://twitter.com/Ling_Lass/status/1335045068701503494
7. Place identity & co-occurrence in Northern Maine:
   https://twitter.com/Katharina_Pabst/status/1335049287097589761
8. The demise of impersonal constructions:
   https://twitter.com/chao_noelia/status/1335052686442553346
9. Determinatives are far from pronouns in English:
   https://twitter.com/brettrey3/status/1335060090941018112
10. Emoji based reactions to the Said Construction:
    https://twitter.com/AliciaStevers/status/1335064270833405952
11. #AboriginalEnglish: BE LIKE, stability and change:
    https://twitter.com/CelesteRLouro/status/1335067954019323909
12. Perception of American English pure vowels:
    https://twitter.com/NaimAfshar/status/1335071401934467074
13. Interactionally situating the power scream:
    https://twitter.com/EMdoesCA/status/1335075479506857986
14. Why are we *quick* to point out, but not *fast*?
    https://twitter.com/demeco_project/status/1335079000155295747
15. What the Italian subjunctive actually means:
    https://twitter.com/salviodigesto/status/1335082822600634368
16. Empirical methods for describing TAM:
    https://twitter.com/AnaKrajinovic1/status/1335086658962780161

17. Positional Preference of Emotion Phrase in Hindi:
    https://twitter.com/spandan_ju/status/1335090484667039746
18. We don't agree (only) upwards:
    https://twitter.com/Andraas/status/1335094108436750337
19. The F2 Robot Interaction System:
    https://twitter.com/f2robot/status/1335097926629126148
20. Acquisition of syntactic negation & NC:
    https://twitter.com/samrinice/status/1335101660633440256
21. Kazym Khanty -en: 2SG possessive—>proprial article:
    https://twitter.com/SK_Mikhailov/status/1335105393861808128
22. Cracking stereotypes the ling of discourse markers:
    https://twitter.com/AichaBelkadi/status/1335109284170969088
23. Lexical classification of Tupí-Guaraní languages:
    https://twitter.com/fthorstensen/status/1335113053571080193
24. The way Spanish and Basque think about causality:
    https://twitter.com/AndreaArioBizar/status/1335116703517331456
25. Social influence on negation in Early Modern Dutch:
    https://twitter.com/leviremijnse/status/1335120535303430144
26. How to clip words in English:
    https://twitter.com/hilpert_martin/status/1335124283526422529
27. The QUD in quantity judgments:
    https://twitter.com/kerbach2/status/1335129528524529664
28. Intensifiers across social media:
    https://twitter.com/tschfflr/status/1335132021216174082
29. Gaze-selection & syntax in multiperson interaction:
    https://twitter.com/cal_virgi/status/1335135674039754754
30. Variation in framing of real-world events:
    https://twitter.com/gossminn/status/1335139354063347713
31. Because-X and because-ellipsis: A comparison:
    https://twitter.com/linguistlaura/status/1335143470684663810
32. who gives what to whom, and how do we know:
    https://twitter.com/evaeva_z/status/1335147032164626433
33. Processing linguistic variation:
    https://twitter.com/rkofreitag/status/1335171551025565700

34. Children use prosody for sentence disambiguation: https://twitter.com/KolbergLeticia/status/1335177147401592832
35. Feedback as the main mechanism of L2 processing: https://twitter.com/AmandaP27090148/status/1335196268042260481
36. Multimodal perception of Brazilian Portuguese: https://twitter.com/Lumamirand/status/1335203546669658112
37. Matrix Language in the Code-Switching in Children: https://twitter.com/KFascinettoZ/status/1335208175130198020
38. Efficient coding: Passive and dative alternations: https://twitter.com/haspelmath/status/1335212273250562048
39. How priming in bilinguals leads to language change: https://twitter.com/EvangeliaAdamou/status/1335218667580252160
40. Bilingual mixed NPs: speech data vs. models: https://twitter.com/MixedNPs/status/1335222452826165249
41. Pragmatic and discursive mechanisms in headlines: https://twitter.com/daniel__pascual/status/1335229942758510592
42. What counts as alternating passive constructions? https://twitter.com/marciamv2/status/1335233794585092097
43. Gricean Secrets: https://twitter.com/anthony69848604/status/1335237503775870983
44. Cats of Twitter: https://twitter.com/BerLinguistin/status/1335241599715074051
45. National language literacy lessons in The Gambia: https://twitter.com/clydeancarno/status/1335245117964300288
46. Ideologies in a university linguistic landscape: https://twitter.com/ruiality/status/1335248929022238720
47. The Decline of V2 in the History of English: https://twitter.com/sophiewhittle95/status/1335260153877237760
48. Translanguaging lens in deaf education: https://twitter.com/AryaneSNogueira/status/1335279092124569601
49. The Changing Language of the Climate Change Debate: https://twitter.com/RDFT58485932/status/1335301693513273347
50. Be that as it may: The Unremarkable Trajectory of the North American English Subjunctive: https://twitter.com/mizlinguist/status/1335309295215390721

51. Teasing on Twitter: An analysis of Donald Trump's tweets:
https://twitter.com/lillapszabo/status/1335312990174896136
52. Frame Semantics & Multimodal Machine Translation:
https://twitter.com/viridiano/status/1335286721269919745
53. One Too Many Plural(s): Taglish Code-Switching:
https://twitter.com/petertorres/status/1335320655143731201
54. A Classless Analysis of Italian Nouns:
https://twitter.com/ulfsbjorninn/status/1335354516275912704
55. Because X: Now I'm an ellipsis and now I'm not:
https://twitter.com/TeapotLinguist/status/1335365957401907200

RiCL Research in Corpus Linguistics

"You have done a great job, but I would make some changes." Concession and politeness in asynchronous online discussion forums

Susana Doval-Suárez – Elsa González-Álvarez
University of Santiago de Compostela / Spain

Abstract – The aim of this study is to provide a preliminary characterisation of concessives in asynchronous online discussion forums and to explore how learners participating in the discussions use concession in combination with other politeness strategies in a collaborative pedagogical context. For this purpose, a corpus of 165 concessive clauses headed by *but* (henceforth, butCs) was extracted from the English component of the *Santiago University Corpus of Discussions in Academic Contexts* (SUNCODAC). First, we explored the co-occurrence of butCs with different lexical features (first and second-person pronouns and adjectives, hedges, boosters and positive and negative sentiment words) which have been reported to be important for this categorisation (Hyland 2005; Musi *et al.* 2018). Then, variations in the frequency of use of these linguistic features were investigated using the Log Likelihood test in relation to different contextual factors: a) message section, b) course period, and c) gender. The results of the quantitative analyses indicate that the typical butC co-occurs with a set of lexical features whose distribution is clearly determined by the discourse function of the two concessive propositions, and by the part of the message in which it appears. Furthermore, the fact that the frequency of all features seems to decrease over time seems to point to an evolution from a more tentative to a more confident tone in posts. The results also confirm the existence of gender-related differences.

Keywords – computer mediated communication; politeness; concessive; mitigation; argumentative

1. INTRODUCTION[1]

1.1 Politeness in computer-mediated communication

The aim of this study is to shed light on the use of argumentative concession in asynchronous online discussion forums. The use of online discussion forums and other types of computer-mediated communication (henceforth CMC) in educational settings has become an extended practice that enables participants to work and construct

[1] We sincerely thank the editors and the two reviewers for taking the time to review the manuscript and providing constructive feedback which improved the original. For generous financial support, we are grateful to the following institutions: 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).

knowledge beyond the time and space constraints of the classroom. In fact, online interaction environments have been reported to be potentially powerful tools for collaborative learning and group communication (Schallert *et al.* 2009; Van Nguyen 2010). As predicted by Jordan *et al.* (2014: 451), CMC has continued to play "a significant role in formal learning as institutions of higher education increasingly offer online and hybrid courses," especially with the challenges brought about by the COVID-19 crisis.

Against this background, we also seek to explore how learners participating in the discussions use concession in combination with other politeness strategies in a collaborative pedagogical context. Therefore, we are interested in issues of face (Goffman 1967) and politeness (Brown and Levinson 1978; 1987). In the list of potentially face-threatening acts (FTAs), Brown and Levinson's theory of politeness includes orders, requests, suggestions, advice, reminders, warnings, offers, promises or criticism (Brown and Levinson 1987: 66-67). These speech acts can be mitigated by using positive and negative politeness strategies, depending on whether they are used to protect positive face (i.e., the universal desire to be appreciated and socially accepted) or to protect negative face (i.e., people's desire to preserve autonomy). Examples of positive politeness strategies include attending to the interlocutor's needs or wants, seeking agreement, softening disagreement, including the writer and the reader in the activity, and showing praise or appreciation, among others. Negative strategies, on the contrary, include being indirect, minimising an imposition, apologising, and impersonalising a situation, among others (Schallert *et al.* 2009: 718).

Even though Brown and Levinson's work has remained influential over the years, it has been frequently challenged. Thus, considerable criticism has come from Watts (1992, 2003), Locher (2004), Locher and Watts (2005, 2008), who argue that Brown and Levinson's model is not "in fact a theory of politeness but rather a theory of facework" that fails to account for "those situations in which face-threat mitigation is not a priority," such as aggressive or impolite behaviour (Locher and Watts 2005: 10). Focusing on the interpersonal dimensions of language used in interaction, they develop the concept of 'relational work', i.e., "the 'work' that individuals invest in negotiating relationships with others" (Locher and Watts 2005: 10). It is important to remark that, in their view, Brown and Levinson's concept of politeness can still be used, but it should be viewed as only a small part of relational work, which, in turn "comprises the entire continuum of verbal

behaviour from direct, impolite, rude or aggressive interaction through to polite interaction" (Locher and Watts 2005: 11).

From the point of view of politeness, the online medium has several peculiarities which inevitably shape CMC interactions. On the one hand, it imposes certain limitations which make participants reinforce the interpersonal links with their partners using markers of affection, interactive responses, and group cohesion expressions (Fernández-Polo and Cal-Varela 2017). On the other hand, the lack of non-verbal clues increases the importance of using politeness to avoid misunderstandings, since FTAs such as "disagreements, criticisms, requests for information or help, and requests for clarification of a prior message" (Schallert *et al.* 2009: 715) are typical of CMC interactions (Herring 2023). This is especially true for those interactions including assessment or evaluation of peers' (L2) writing (Cal-Varela and Fernández-Polo 2019; Pyo and Lee 2019), as is the case with the discussion forums in this study (cf. section 2). In these language learning contexts, where the emerging virtual communities have been found to promote interaction and diminish anxiety of communication (Deris *et al.* 2015: 79), the presence of FTAs also leads participants to soften their comments through mitigation strategies.

The emerging interest in politeness issues in CMC has produced a substantial body of research. Different CMC modes have been covered in the literature: e-mails (Harrison 2000; Vinagre 2008), *Wiki* exchanges (Li 2012), blogs (Puschmann 2010), and synchronous and asynchronous discussion forums (Herring 1994; Park 2008; Schallert *et al.* 2009), among others. In general, positive politeness strategies have been found to be more frequent than negative strategies in CMC. This is often attributed to the participants' need to create solidarity (Park 2008; Vinagre 2008) and to maintain accuracy, while avoiding the ambiguity and indirectness that is often brought about by negative politeness (Morand and Ocker 2003). However, negative politeness seems to be more frequent in CMC than in face-to-face interaction (Carlo and Yoo 2007).

One of the topics that has attracted the most interest is gender differences in politeness. Thus, Herring (1994) reports "a tendency for women to favour positive politeness and men negative politeness," although the most remarkable difference she finds is that flaming (i.e., posting angry or insulting messages) is "practised almost exclusively by men" (Herring 1994: 291). Similar conclusions are reached by Hall (1996) and Herring (1996; 2000), who also suggest that, while women tend to be more worried about politeness, men tend to engage in more FTAs and "to be more concerned about

threats to freedom of expression than with attending to others' social 'face'" (Herring 2000: 3). Similarly, Guiller and Durndell (2006) found that, in educational forums, males tend to use more authoritative language and argumentation than females. However, Herring (1996) also suggests that these gender differences may disappear in mixed-group forums where members of the minority gender tend to imitate the majority gender communicative style. Likewise, Savicki *et al.* (1996) and Tet Mei *et al.* (2023) show that CMC is gradually becoming more gender-neutral in terms of politeness features, possibly because participants tend to accommodate each other's gendered language styles (Thomson and Murachver 2001).

Assuming the existence of gender differences in studies of language use is, however, controversial. In fact, the pre-conception that women and men can be viewed as internally homogeneous groups has been progressively abandoned in the feminist literature (Cameron 1992). According to 'the dynamic approach' to gender (West and Zimmerman 1987; Crawford 1995) gender is not "a static, add-on characteristic of speakers, but is something that is accomplished in talk every time we speak" (Coates 2004: 7). In addition, exploring gender differences in the context of CMC research is criticised on the grounds that CMC possesses a "degree of anonymity that makes the gender of online communicators irrelevant or invisible" (Graddol and Swann 1989, as cited in Herring and Stoerger 2014: 567). In contrast, Yates (2001) argues that the gender differences found in face-to-face research are sometimes magnified in CMC, since "gender is often visible in CMC on the basis of features of a participant's discourse style" (Guiller and Durndell 2006: 368). Additionally, Herring and Stoerger (2014: 576) remark that most instances of asynchronous CMC are not anonymous and, even when pseudonyms are used, gender can still be identified since "communicators give off cues through their interactional style and message content."

Other issues dealt with in the CMC literature on politeness include differences in politeness in CMC versus non-CMC discourse (Brysbaert and Lahousse 2019), the relationship between politeness and discourse functions (Schallert *et al.* 2009), and the effects of time in the communicative and politeness practices of online learning communities, and L1-related differences in strategy choice (Fernández-Polo and Cal-Varela 2017; Cal-Varela and Fernández-Polo 2020).

1.2. Argumentative concessives and politeness

Despite the important argumentative value ascribed to concessive connectors in the literature on academic discourse (Biber *et al.* 1999; Couper-Kuhlen and Thompson 2000), little research has been conducted on the role played by these rhetorical relations in CMC. A few exceptions can be found. Tanskanen and Karhukorpi (2008), for instance, explore how participants in e-mail conversations use concessives to correct themselves. Their study suggests that when participants use concessives to repair claims that may cause disagreement, they are adopting the perspective of their fellow communicators and negotiating affiliation "in a dialogical manner" (Tanskanen and Karhukorpi 2008: 1587). Drawing on an interest in online forums as channels for public dialogue on current political and social issues, Swanson *et al.* (2015) deal with concession in the context of argument mining. Thus, they suggest that statements containing "specification, contrast, concession and contingency markers are more likely to contain good argumentative segments" (Swanson *et al.* 2015: 218). Most remarkably, concessives are the focus of a study conducted by Musi *et al.* (2018), who test the hypothesis that argumentative concessions can be used as persuasive strategies by calculating their frequency in persuasive vs. non-persuasive discourse. For this purpose, they use a CMC dataset, the *ChangeMyView Subreddit* platform, "where multiple users negotiate opinions on a certain issue willing to change their point of view through other users' arguments" (Musi *et al.* 2018: 2). Although their results suggest that concessions do not make the arguments more convincing in this specific context, they argue that this is because their persuasive value is "context-bounded and crucially depends on the rhetorical situation" (Musi *et al.* 2018: 16).

Outside the CMC context, the literature on concessives has referred indirectly to their role as politeness strategies. Thus, Biber's (1988) multidimensional approach associates concessives with other mitigating devices such as hedges or downtoners; in this model, concession is a marker of non-assertiveness, since it indicates the possibility that other options are true (Monaco 2017: 138). Furthermore, it is often mentioned that concessives are used to increase the hearer's positive attitude towards the speaker's opinion (Mann and Thompson 1988), since "recognizing the validity of the hearer's standpoint before expressing disagreement can avoid FTAs acts and is perceived as reasonable by the hearer" (Couper-Kuhlen and Thompson 2000: 381). Additionally, some studies have emphasised the correlation between the use of concessive connectors and the

presence of opinion, evaluation, and argumentation (Swanson *et al.* 2015).

The notion of 'concession' used here is based on Couper-Kuhlen and Thompson's (2000: 381) definition of concessives as three-part sequences in which: 1) the first speaker makes a point (X), 2) the second speaker concedes the validity of this point (X'), and 3) the second speaker makes a potentially contrasting point (Y). This description provides the basis for Musi *et al.*'s definition of 'argumentative concessives' (ACs) as a type of concessive in which "the proposition introduced by the connective – B –, which denies the expectations brought about by a preceding proposition, expresses the speaker's standpoint" (Musi *et al.* 2018: 5). According to these authors, at a semantic level, the conceding proposition (or proposition A) of ACs typically includes agreement or a positive evaluation of the statement previously presented by the other speaker, while the denial-of expectations proposition (or proposition B) tends to include (mitigated) criticism. Additionally, Musi *et al.* (2018) suggest that ACs can be characterised by referring to the linguistic features that tend to co-occur with them. Their list of features includes: a) hedges (defined as lexical and syntactic means of decreasing the writer's responsibility "for the extent and the truth-value of propositions and claims, displaying hesitation, uncertainty, indirectness, and/or politeness to reduce the imposition on the reader" (Hinkel 2005: 30)); b) positive and negative sentiment words (since ACs usually contain opinion on the other posts); c) first and second personal pronouns and adjectives (since ACs "dialogically point to the stance taken by the previous speaker" (Musi *et al.* 2018: 10)); and d) modal verbs, which indicate that what is expressed in proposition B is not 'unassailable.'

1. 3. The current study

This project aims to provide a preliminary characterisation of concessives in the *Santiago University Corpus of Discussions in Academic Contexts* (SUNCODAC 2021) and to show the relevance of politeness for this characterisation. More specifically, we are interested in exploring how L1-Spanish EFL learners participating in this discussion forum use *but*-concessives (henceforth, butCs) for argumentation.

The decision to include only butCs was motivated by the fact that these connectives have been found to be the most frequent concessive marker in different discourse types (Grote *et al.* 1997; Izutsu 2008; Taboada and Gómez-González 2012; Gómez-González

2017).[2] Additionally, *but* represents 85 per cent of concessive markers in discussion forums (Musi *et al.* 2018) and 52 per cent of all concessive markers in the English component of SUNCODAC (Doval-Suárez and González-Álvarez 2021). Barth (2000: 418) explains that the reasons for this prevalence of *but* are not only connected with the fact that they are paratactic constructions which "facilitate on-line production," but also, and most importantly, with the fact that they "provide an opportunity for face work by leaving the speaker room to manoeuvre and by attending to the recipient's need for politeness." Additionally, Uzelgun *et al.* (2015) suggest that the *yes … but*-construction plays a key role in the study of (dis)agreement space by presenting what is accepted as opposed to what is criticised.

Drawing on Musi *et al.*'s (2018) characterisation of concession, we focused on the use of butCs in combination with hedges, positive and negative words, and first and second personal pronouns and adjectives.[3] Additionally, boosters were also included as a category in this characterisation. The reason for this is that, together with hedges, boosters can function as stance markers or markers of epistemic modality, since they are used by a speaker/writer "to signal different degrees of certainty concerning the validity of the information" and "to increase or decrease the illocutionary force of speech acts" (Holmes 1982: 11). Therefore, boosters and hedges are two sides of the same coin.

2. METHOD

2.1. Participants and data source

SUNCODAC, the corpus used in this study consists of student forum discussions gathered over a span of four years at the University of Santiago de Compostela (USC).[4] These discussions were an integral part of an English-to-Spanish translation course designed for second-year undergraduates, primarily majoring in English at USC. The forum served as a supplementary tool alongside traditional face-to-face teaching, and students actively contributed at three distinct times during the semester: the beginning (period 1), middle (period 2), and end (period 3).

[2] The concessive value of *but* has been generally ignored in the literature. For a detailed description of the concessive, contrastive, and corrective meanings of *but*, see Izutsu (2008).
[3] These authors also include modals as a separate category, but our study focused only on modals working as hedges.
[4] http://www.suncodac.com/

As shown in Table 1, the corpus contains a representation of English, Spanish, and Galician used as first (L1) and second (L2) languages by students of different nationalities. The subjects are L1 and L2 English speakers of several L1 backgrounds, mainly Spanish, Galician, English and Chinese, but this study concentrates on L1-Spanish participants' productions in L2 English.

Languages	**Words**	**Posts**		**Number of participants**				
Spanish	232,440	1,521	Gender	L1 Sp./Gal.	L1 English	L1 Chinese	L1 Other	Total
Galician	18,547	119	Female	295	17	56	30	398
English	328,537	1,724	Male	87	8	20	7	122
Total	579,524	3,364	Total	382	25	76	37	520

Table 1: A description of SUNCODAC

A detailed description of the activity can be found in Cal-Varela and Fernández-Polo (2020: 46–47). Every week, a practical session was allocated for in-class discussions on a translation topic, followed by an online discussion. To facilitate this, distinct weekly forums were created within the Moodle platform. Each forum was overseen by a student who was assigned the role of moderator. The activity unfolded through five stages:

1) Lecturers' instructions. A single opening post by the lecturers including the source text, the moderator's name, basic instructions, and deadlines.
2) Moderator's first translation.
3) Peer feedback. This is the core of the discussion and consists of messages where the moderator's classmates make comments and suggestions for improvement and discuss the suitability of different translation solutions.
4) Moderator's improved version and summary of discussion.
5) Instructor's assessment and appraisal of the activity.

It should be noted that most of the corpus consists of feedback messages, that is, posts belonging to stage 3. Therefore, posts from this stage are the central part of the discussion and the bulk of the corpus. Each of these feedback texts may have different sections or moves (Fernández-Polo and Cal-Varela 2018):

1) Pre-proposal: provides an overall evaluation of the translation, may touch upon potential weak points, mention other aspects like task difficulty, or include expressions of congratulations.
2) Proposal: represents the core of the message, listing problems in the translation

provided and offering suggestions for improvement.

3) Post-proposal: is often quite similar to the pre-proposal (but appears less frequently).

4) Opening and Closing sections: these two sections exhibit an epistolary style. The opening section features a salutation, and the closing section includes various expressions of farewell.

The different sections in peer feedback posts are illustrated in (1). [5]

(1) **OPENING**
Hi everyone!

PRE-PROPOSAL
I think that your translation is very good, but I would change a couple of things.

PROPOSAL
For example, instead of "porque afecta a la recuperación de las heridas." I put "ya que afecta a la recuperación de lesiones" because I think that it refers to a general term (lesiones). Then, in "Este líquido necesita ser reemplazado rápidamente para contribuir a la recuperación de las articulaciones doloridas y de los músculos" I put "Este fluído debe ser reemplazado rápidamente para eliminar los dolores en las articulaciones y músculos" because it sounds more natural, more like a colloquial language.

POST-PROPOSAL
For the rest my translation is the same as yours, so that's all.

CLOSING
Regards!

(16MPU_ The best food for footballers 2016-A)

2.2. Procedure

Since this is a small-scale study, the first step was to create a subcorpus of butCs. Therefore, using the corpus search tool, a sample of 165 butCs produced by the L1-Spanish group was extracted from the English component of SUNCODAC feedback messages This sample represents 15 per cent of the overall occurrence of this marker in the whole corpus. The butCs are uniformly distributed across sections, gender groups and periods, i.e., we selected equal numbers of butCs for each level of the different variables used as corpus design criteria: gender, post section, and post period.

[5] All examples included in the article are corpus examples which have not been altered. This means they may include spelling errors and typos, among other types of mistakes.

The creation of the subcorpus was followed by the automatic extraction of examples containing the different lexical features under study using *Wordsmith Tools 7* (Scott 2016), and by the manual disambiguation of examples. The list of lexical features was constructed by referring to previous studies. Thus, we used the lists of hedges found in Hyland (2005), the list of intensifiers used by Hinkel (2005), and, in order to select the positive and negative sentiment words, we chose the sentiment/opinion lexicon published by Hu and Liu (2004), also adopted by Musi *et al.* (2018).

The variables considered in the subsequent quantitative analyses were the concessive proposition (A/B), the post section, course period, and gender.[6] The quantitative analyses used Log Likelihood to test for statistically significant differences.

2.3. Research questions

In order to describe how a specific type of concessive (i.e., butC) is used in combination with other politeness strategies in a specific CMC mode (i.e., online discussion forums), our study addresses the following five research questions:

1) How frequently do butCs co-occur with the following lexical features: boosters, hedges, positive and negative sentiment words, and first and second personal pronouns and adjectives?

2) What is the distribution of these lexical features in propositions A and B of the concessive?

3) Are there any significant differences in the frequency and distribution of these lexical features between message sections (preproposal/proposal)?[7]

4) Are there any significant differences in the frequency and distribution of these linguistic features between butCs produced at the beginning and the end of the term (i.e., period 1 and period 3)?

5) Are there any significant differences in the frequency and distribution of these linguistic features between butCs produced by male and female participants?

[6] Although we are aware of the problematic status the category 'gender' (cf. Section 1), we will stick to the two-way ('masculine' vs. 'feminine') classification of the participants' gender made by the SUNCODAC compilers.

[7] Post-proposals are not considered here because no examples of butCs were found in this section.

3. RESULTS AND DISCUSSION

3.1. Towards a characterisation of concessives in SUNCODAC

The first step in the characterisation of concessives involved checking whether butCs in SUNCODAC followed the interactional and semantic patterns described by Couper-Kuhlen and Thompson (2000: 38). Our analysis revealed that most butCs in SUNCODAC typically form part of a tripartite sequence in which: a) Student 1 posts a translation, i.e., makes a point (X); b) in another post, Student 2 concedes the validity of the other student's point in proposition A (the conceding move) by means of partial agreement, approval or praise for the proposed translation (X'); and c) Student 2 goes on to make a potentially contrasting point in proposition B (the denial of expectation move) by suggesting changes to the original translation (Y). Additionally, and drawing on Musi *et al.* (2018)'s semantic characterisation, our butCs were found to consist of a conceding move containing positive sentiment or agreement in proposition A and a denial-of-expectation move containing some sort of mitigated criticism or imposition in proposition B (cf. Figure 1).

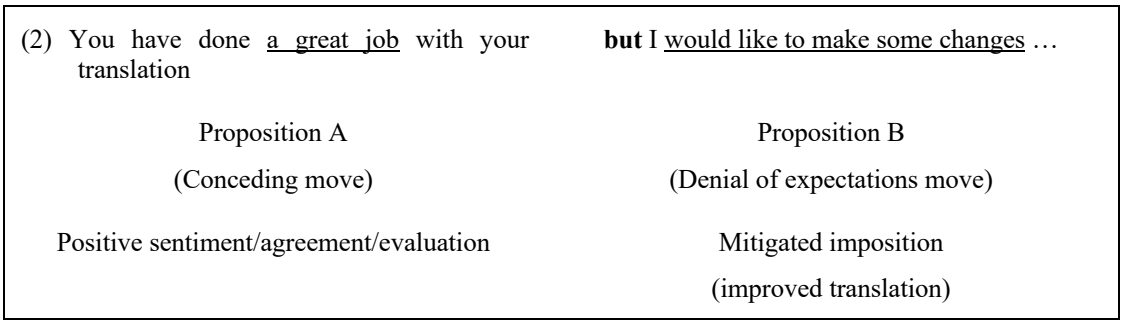

Figure 1: Typical concessive pattern (i.e., pattern 1) in SUNCODAC

However, the analysis revealed that this pattern (henceforth, pattern 1), though prevalent in the corpus, could not account for all the instances of butCs. Thus, a corpus-based approach was adopted to detect other interactive/semantic patterns. As a result of the manual analysis of concordance lines, two additional patterns emerged, whose respective frequencies are shown in Table 2.

	Number	**Percentage**
Pattern 1	124	75.2
Pattern 2	19	11.5
Pattern 3	22	13.3
Total	**165**	**100.0**

Table 2: Concessive patterns in SUNCODAC butCs

Figure 2 shows that, in pattern 2, which represented 11.5 per cent of the instances of butCs, the order was occasionally reversed so that proposition A was the one including the alternative translation, while proposition B was the one containing positive evaluation:

(3) And I chose "James R . Flynn descubrió que" instead of "reparó"	**but** I think the verb you chose works just as well (16ASE_Intelligence_2016-B)
(4) "Es cierto que ..." to me it sounds better,	**but** as I said yours still makes sense. (16ASE_Shrinking families_2016-A)
Proposition A	Proposition B
Alternative translation	Positive evaluation/agreement

Figure 2: Alternative concessive pattern in SUNCODAC

Finally, a miscellaneous pattern (pattern 3) was also identified to account for variations of the preceding two patterns as in (5), where proposition A includes the alternative translation and B is an evaluation of this alternative. Another example is (6) where a butC appears inside another concessive headed by *although.* This heterogeneous pattern 3 represents 13.3 per cent of the total tokens of butCs.

(5) "Finally, it sounds better for me "largas mensulas piramidales invertidas", but maybe it is a bit stiff." (16DRP_Male_The gift of the gab_2016A)

(6) "Although this is a good translation, I would use "intentar" instead of "tratar", but it is just because it sounds more casual for me." (17AGO_The river_2017-A)

The final step followed in order to describe concessives in our corpus involved exploring the potential co-occurrence of butCs with boosters, hedges, positive and negative sentiment words, and first and second personal pronouns and adjectives (henceforth, *I*-words, *you*-words and *we*-words). In order to determine the importance of these elements for their characterisation, two measures were used: a) the proportion of butCs including each of these features (cf. Table 3), and b) their distribution in propositions A and B (Table 4).

	Number of concessives	Percentage of concessives
I-words	145	87.9
Positive	106	64.2
You-words	93	56.4
Hedges	85	51.5
Boosters	60	36.4
Negative	19	11.5
We-words	15	9.1
Zero features	19	11.5

Table 3: Frequency of butCs containing at least one token of each of the selected linguistic features

The data in Table 3 show that only 11.5 per cent of butCs in our corpus exhibit no examples of the linguistic features under consideration. As for the butCs containing at least one token of each of these features, their frequencies are presented in decreasing order: 87.9 per cent of the butCs contain at least one *I*-word, which highlights the importance of first-person voice. A similar incidence of 'egocentric deictic reference' has been detected in other forms of CMC, such as corporate blogs (Puschman 2010: 181), where participants are likely to feature prominently in their own discourse. In contrast, butCs with *we*-words are placed much lower in the rank (9.09%), but are also noteworthy, since they are sometimes used as a positive politeness strategy with the purpose of including the writer and the reader in the activity, thus reinforcing the sense of community of learning (7), an effect that can be also achieved by using a combination of *I*- and *you*-words, as in (8). In other cases, *we*-words are simply used as an instance of generalisation (9) or as negative politeness strategies to impersonalise an imposition, as in (10):

(7) May I start saying that is a great translation, but I completely agree with the suggestions **our** mates made, like 17DVM 's, to make the text more natural (17ARB_The river_2017A)

(8) **I** like **you**, wrote that the little door behind the curtain was 40 cm tall, but I read some of our classmates answers and I have to agree with the ones that put instead something in the lines of "de dos palmos de altura" or something of the sort. (16ASE_Alice in Wonderland_2016-A)

(9) From my point **we** can consider your translation as more "technique" in the sense that you are using specific lexicon, but here **we** are writing in a newspaper and the most important thing is to arrive to the greatest possible number of people. (17AHF_Smart jacket_2017)

(10) I think that translating "who" by "los cuales" does not sounds too formal, it's a relative more complex than "que" but in this context **we** can use "que", "los cuales" or "quienes" because they tree have the same meaning in here. (16ACC_ English on the march_2016-B)

The percentage of concessives including instances of second person reference is smaller but still important (56.4%) in a context characterised by appeals to other users. Furthermore, if *I*-words, *you*-words and *we*-words are considered together, their high prevalence may point to the dialogical character of discussion forums, which, like other types of CMC are said to "bristle with first and second person pronouns" (Jonsson 2015: 215).

Leaving out pronouns, the most characteristic feature of ACs seems to be the presence of at least one positive word or a hedge in more than half the examples. On the one hand, positive words, which appear in 64.2 per cent of the concessives, are used for subjective evaluation and are an indication of the presence of positive politeness strategies such as praise, appreciation, or gratitude (Schallert *et al.* 2009: 718). On the other hand, the appearance of hedges in 51.5 per cent of the examples may suggest a cautious, non-assertive kind of discourse. Hedges in our corpus are typically used as negative politeness strategies, i.e., in order to minimise the imposition represented by the suggested changes to the original translation (11).

(11) Hello! I agree with you, 16VVE, but I **would** change the translation. (16AFF_Emergency_2016-A)

Finally, the high incidence of hedges and positive words contrasts with the relatively low percentage of butCs containing boosters and negative words, both of which fall considerably below the halfway point (with percentages of 36.4% and 11.5% respectively). A tentative explanation may be that boosters often help participants to construct a more authoritative or confident kind of discourse, which conflicts with the attenuation effect of hedges. Furthermore, the predominant function of the examined concessive clauses seems to be mitigation rather than the overt expression of criticism by means of negative sentiment. In fact, as could be seen in (5), repeated as (12) for convenience, and (13) below, many of the negative words in the corpus are not used to criticise the other participant's translation, but to evaluate the speaker's own proposal, in an attempt to diminish the FTA of imposing a change:

(12) Finally, it sounds better for me "largas mensulas piramidales invertidas", but maybe it is a bit **stiff**. (16DRP_The gift of the gab_2016-A)

(13) It sounds too much formal and non-natural for me the first time I read it. So sorry. But doesn't matter, it's just my **stupid** opinion. (16DRP_Emergency_2016-B)

The final step in the characterisation of ACs involved exploring the overall frequency of each of these features in the *but*-corpus and their distribution in the two concessive propositions (A and B). For this purpose, we calculated their raw and relative frequencies, as shown in Table 4.

	Overall		**Proposition A**		**Proposition B**		**Log Likelihood (LL)**
	Raw	**Fpttw**	**Raw**	**Fpttw**	**Raw**	**Fpttw**	
I-words	273	453.5	127	466.2	146	443.0	+0.18
You-words	129	214.3	95	348.8	34	103.2	+42.83**
Positive	127	211.0	94	345.1	33	121.2	+43.32**
Hedges	101	167.8	24	88.1	77	282.7	-20.07**
Boosters	70	116.3	48	68.6	22	31.4	+15.48**
Negative	20	33.2	9	3.0	11	40.4	-0.00
We-words	17	28.2	7	25.7	10	30.3	-0.11

Table 4: Distribution of hedges, negative and positive sentiment words in the two concessive propositions

The results of the Log Likelihood test indicate the existence of statistically significant differences (**) between the two concessive propositions regarding the frequency of *you*-words, positive sentiment words, hedges and boosters. On the one hand, *you*-words and positive words are significantly more frequent in proposition A (LL=+43.32; p<.05 and LL=+42.83; p<.05, respectively), since this is the one usually containing some sort of (boosted) praise of the other participant's translation. An illustration of the occurrence of *you*-words and positive sentiment in proposition A can be found in (14) and (15) below:[8]

(14) Hi everybody! **I think that <u>your</u> translation, 16MSF, is <u>excellent</u>** but I have some differences (16ASP_Alice in Wonderland_2016-A).

(15) First of all, congratulations to you 16NBA, **<u>you</u> have done a <u>wonderful</u> job translating this text**, but I would like to point out some things that I translated differently. (16AEG_ The best food for footballers_2016-A)

On the other hand, our findings indicate that hedges are overused in proposition B (LL=-20.07; p<.05), which confirms the tendency observed by Musi *et al.* (2018), who refer to the frequent presence of modal verbs, a specific type of hedge, in proposition B of argumentative concessions. This is coherent with the fact that proposition B is the one presenting the improvements to the original translation made by the other student. In line with Hyland (1996), the presence of hedges might serve to make this proposal easier to accept by softening its tone. This can be seen in examples (16) and (17):

(16) Hi 16VPL ! You have done a great translation, **but <u>I think</u> <u>some</u> things <u>could</u>**

[8] In all the examples that follow, the proposition under discussion appears in bold type and the co-occurring linguistic feature is underlined.

be changed (16AFF_Shrinking families_2016-A).

(17) I think you have done a great job with your translation **but I have some suggestions that perhaps, they may help you (or not)**" (17AAR_Cellscope Oto_2017-B)

As for negative words, the results are inconclusive, since no statistically significant differences were found between the two propositions. In other words, although negative sentiment is slightly more frequent in proposition B (18), it seems that its presence in proposition A is not necessarily connected with the evaluation of the translation under discussion. For instance, the word *complicated* in (19) is not used to qualify the other student's translation, but to highlight that they did a good job despite the difficulty of the task:

(18) I understand that you are trying to keep the original tone of the text, but as the others said, **it sounds weird here**. (16DRP_The best food for footballers_2016-A)

(19) **Firstly, I wanted to say I think you did a good job translating this extract of the text since I found it a bit complicated**, but there are some words and expressions I'd change. (17AGG_The river_2017-A)

We have, therefore, established a tentative characterisation of concessives in our corpus in terms of some of the linguistic features that co-occur with them. The following sections will explore if these findings can be further qualified by considering (contextual) factors such as course period, message section, and participant's gender.

3.2. Frequency and distribution of lexical features in butC in relation to different contextual factors

3.2.1. Message sections

We have already described the posts as consisting of different sections or moves. According to the corpus compilers (Fernández-Polo and Cal-Varela 2018), forum posts in SUNCODAC can be categorised as a genre, understood as comprising standard sequences of moves, or text segments that play identifiable roles within the overall structure. The authors found that the structural components observed in SUNCODAC exhibit move-like properties, characterised by distinctive language and specific text positions (Fernández-Polo and Cal-Varela 2018: 192).

To determine whether the two sections containing butCs in our subcorpus also have distinctive language characteristics, we conducted a comparison of the frequency of the

features considered for our study between the preproposal and proposal sections. The results are shown in Table 5.

	Proposal		**Proposal**		**Log Likelihood (LL)**
	Raw	**Fpttw**	**Raw**	**Fpttw**	
Hedges	60	**350.7**	41	156.8	**+16.16****
Boosters	14	81.8	56	**214.1**	**-12.29 ****
Negative	2	11.7	18	**68.8**	**-8.83****
Positive	60	**350.7**	71	271.5	**+2.11**
I-words	149	**935.1**	126	443.6	**+23.96****
You-words	92	**420.8**	37	221.8	**+53.31****
We-words	8	**46.8**	9	34.4	**+0.39**

Table 5: Frequency and distribution of lexical features of butCs appearing in different message sections

As shown in Table 5, the Log Likelihood test reveals a significantly higher frequency of hedges, *I*- and *you*-words in preproposals, but a significantly lower frequency of boosters and negative sentiment words. No significant differences were observed in the frequency of use of positive words and *we*-words.

The fact that hedges in butCs are significantly more frequent in preproposals than in proposals can be explained by the fact that preproposals in SUNCODAC are often used to announce criticism, and hedges often co-occur with critical comments, precisely because of their ability to keep the distance between what is being said and the actual writer's opinion. Thus, any conflicts that could arise from explicit claims to an absolute truth are avoided. Similarly, Cal-Varela and Fernández-Polo (2020) identified hedges as part of the mitigating strategies of preproposals in the Spanish subcomponent of SUNCODAC.

While hedges are more frequent in preproposals**,** boosters and negative words are significantly more frequent in proposals, whereas positive words and *we*-words appear with similar frequencies in both types of sections. A tentative interpretation of these results is that in proposals the focus is on criticism. Thus, boosters are used for two apparently contradictory purposes: to boost criticism, qualifying the writer's commitment to the truth of the proposition; and to project added politeness, sincerity and truthfulness. As for negative words, it has already been observed (cf. Section 3.1.) that they are sometimes used to qualify the speaker's own proposal, serving as negative politeness strategies that reduce the force of the imposition caused by suggesting an alternative translation.

On the other hand, positive words are slightly overused in proposals, which is in

line with Tan *et al.* (2016), who state that persuasive opening arguments (which could be the equivalent of SUNCODAC preproposals) use fewer positive words, suggesting more complex patterns of positive emotion in longer arguments appearing later in the message (i.e., proposals). However, this finding should be taken with care since this overuse is not statistically significant.

3.2.2. Course period

It has been observed that the presence of politeness devices, which serve to mitigate critical comments and contribute to a more congenial learning environment, might be anticipated to evolve over the duration of the course, particularly if the group develops into a genuine community of inquiry (Fernández-Polo and Cal-Varela 2017: 260). To investigate whether this evolution occurs in the forum discussions being analysed, we conducted a comparison of the frequency of different features between Period 1 and Period 3. The results can be seen in Table 6.

	Period 1		**Period 3**		**Log Likelihood (LL)**
	Raw	**Fpttw**	**Raw**	**Fpttw**	
Hedges	51	324.8	30	207.0	**+ 3.95 ****
Boosters	19	121.0	28	193.2	- 2.53
Negative	1	6.4	1	6.9	- 0.00
Positive	46	293.0	19	131.1	**+ 9.50 ****
I-words	109	694.3	79	545.2	+ 2.70
You-words	38	242.0	49	338.2	- 2.42
We-words	6	38.2	7	48.3	- 0.18

Table 6: Frequency and distribution of lexical features of butCs in first and last course periods

Table 6 shows a significant decrease in the use of hedges and positive sentiment words over the time span of the course, which can indicate that as participants get to know each other, they feel less need to mitigate the force of the criticism. On the contrary, although not significant, there is a decrease in the use of *I*-words, and an increase in the frequency of *you*- and *we*-words, which could point to an evolution from a "mostly monologic, informational and author-centred" kind of post to "a progressively longer post with […] a heightened awareness of the dialogic and multi-party nature of the exchanges" (Fernández Polo and Cal-Varela 2017: 256).

An evolution over the time span of the course, but pointing in the opposite direction, was observed in previous studies on the use of mitigation strategies in CMC,

where intensity by repetition of the same strategy and attenuation effort (measured by the combination of different attenuating strategies) seem to increase with time, "suggesting that students become increasingly aware of the need to step up interpersonal work" (Cal-Varela and Fernández-Polo 2020: 50).

3.2.3. Gender

Contradictory findings have been obtained in previous studies of the influence of the gender factor on the use of mitigation strategies in discussion forums (cf. Section 1.1). Table 7 shows the distribution of features in relation to gender in the present study.

	Male participants		**Female participants**		**Log Likelihood (LL)**
	Raw	**Fpttw**	**Raw**	**Fpttw**	
Hedges	30	154.1	71	**298.4**	**- 9.93****
Intensifiers	16	80.1	54	**227.0**	**-15.67****
Negative	18	**90.1**	2	8.4	**+17.68****
Positive	76	**390.3**	55	231.2	**+ 8.9 ****
I-words	97	498.2	179	**752.4**	**- 12.49****
You-words	32	164.4	98	**411.9**	**-24.56****
We-words	3	15.4	14	**58.9**	**-5.93****

Table 7: Frequency and distribution of hedges and positive words in butCs produced by male and female participants

The Log Likelihood tests show statistically significant differences between male and female participants. Hedges, boosters, and pronouns are significantly over-represented in females' posts, whereas positive and negative sentiment words are significantly over-represented in posts produced by male participants. The significantly higher frequency of hedges in women's posts seems to confirm assumptions that females tend to use more attenuated speech forms (cf. Guiller and Durndell 2006; Hall 1996; Herring 1994; 1996; 2000), since these features can be used both to attenuate criticism (positive politeness) or to attenuate imposition (negative politeness). Additionally, the higher frequency of personal pronouns in females' posts could be interpreted as an indication of a more dialogical kind of discourse. Finally, considering that the connotations of boosters can vary based on the words they modify, a more in-depth qualitative investigation is necessary for a nuanced interpretation of the findings.

4. CONCLUSIONS

As shown in the previous sections, addressing the description of concessives by referring to their role as politeness strategies is especially relevant for the study of CMC contexts which include assessment and evaluation of peers' writing. The characterisation of butCs carried out in this study was approached in several steps. First, we identified the interactive patterns that are typically followed by butCs and concluded that they usually stick to the structures previously mentioned (Couper-Kuhlen and Thompson 2000; Musi *et al.* 2018). This means that proposition A is semantically characterised by the presence of praise or agreement with the other student's translation, and proposition B is semantically characterised by the presence of mitigated imposition. However, other patterns emerged from the study, which call for a more detailed analysis including more examples and other concessive connectors.

Then, we explored the relative importance of the co-occurrence of butCs with first and second personal pronouns and adjectives, hedges, boosters, and positive and negative words by considering the frequency of this co-occurrence. The fact that only a small percentage of butCs contains no instances of these features seems to indicate that their presence is highly relevant in this characterisation. Furthermore, the high overall incidence of butCs containing *I*- and *you*-words points to a type of discourse in which the high prevalence of first-person voice combines with the importance of the appeal to other users, as happens in texts of a dialogical nature. Also, the abundance of butCs with positive words and hedges suggests that participants in these discussions are focused on "phrasing things in such a way as to take into consideration the feelings of others" (Morand and Ocker 2003: 2). This concern for politeness becomes particularly important in a context where the interactions typically involve assessing each other's production.

Additionally, our findings reveal that the typical distribution of these lexical features in SUNCODAC butCs is clearly determined by the proposition, and that this distribution directly mirrors the semantic and interactional function of each proposition. Thus, boosters, *you*-words and positive sentiment words feature prominently in proposition A, while proposition B is clearly marked by the presence of hedges. In contrast, no statistical differences were found in the case of negative words and boosters, whose low frequency may be connected with the fact that SUNCODAC participants tend to avoid overt criticism of the other participants' translations (i.e., they try to minimise threats to positive face), and also avoid presenting their alternative translations in a way

that can be perceived as a threat to their classmates' negative face (hence the occasional use of negative words to qualify their own suggestions for improvement).

We contend that the emerging characterisation is relevant for Brown and Levinson's model of politeness for three reasons: a) proposition B typically contains a FTA, i.e., an imposition realised as a suggestion for improvement of another student's translation, b) this imposition is typically mitigated by means of hedging, an example of the workings of negative politeness, and c) the FTA in proposition B is typically preceded in proposition A by some sort of positive politeness realised as positive evaluation or agreement with the other student. Furthermore, this characterisation may afford a new insight into the use of politeness strategies not only in asynchronous online discussion forums but also in other CMC modes as well.

In order to address research questions 3, 4 and 5, we investigated if the previous characterisation could be further qualified by considering two task-related factors (message-section and course period) and one participant-related factor (gender). The attested variations in the frequency of occurrence of the different lexical features in the two message sections indicate that these features can be used to characterise preproposals and proposals as distinct moves. Again, the fact that hedges and *I*- and *you*-words are significantly overused in preproposals, while boosters and negative words are significantly overused in proposals, mirrors the functions each section has in the post, which is in line with results obtained in previous studies (Fernández-Polo and Cal-Varela 2018). As for the effects of time, our results clearly point to a significant decrease in the use of hedges and positive words over time, which might imply that as the term progresses, participants are less concerned about politeness issues. We also traced an increase in the use of *you*- and *we*-words which could be a symptom of a gradual evolution towards a more dialogic and multi-party type of discourse.

The existence of gender-based differences in the use of politeness strategies is by no means uncontroversial, and our findings regarding this issue are rather inconclusive. On the one hand, the existence of significant differences between male and female participants in the frequency of the different lexical features might suggest that they have different styles. Thus, our results show that females significantly favour hedges, boosters and pronouns, while men favour the expression of both positive and negative sentiment. In terms of politeness strategies, women tend to mitigate more while men seem to praise more, but also to impose or criticise more often. On the other hand, the overrepresentation of hedges in

posts written by female participants (which suggests that they use a more attenuated style) seems to conflict with the fact that they also overuse boosters (an indication of assertiveness), which points to the need to adopt a different perspective in the study of gender and politeness. However, given the small size of the sample used in the study, these results need to be taken with care, and should be tested on a larger and more representative number of examples, which calls for a larger-scale study with a more balanced representation of male and female participants.

While awaiting the bigger picture, we have brought forward a preliminary characterisation of butCs in SUNCODAC, with some features yielding a neater description than others and some variables clearly being more significant than others. Future analysis should reveal the extent to which this characterisation can be extended to other types of concessives. In addition, further research will necessarily involve a refinement of the lists of lexical features which are relevant for the characterisation of concession. All in all, we have described how concession and other politeness strategies work together towards "creating a comfort zone in which to exchange ideas as well as motivating students' participation" (Schallert *et al.* 2009: 715) in the discussion and, hence, in the learning process. We believe that our study has contributed to a better understanding of the role of this rhetorical relation in discussion forums, but its role in other types of CMC still needs to be investigated.

5. REFERENCES

Barth, Dagmar. 2000. That's true, although not really, but still: Expressing concession in spoken English. In Elizabeth Couper-Kuhlen and Bernd Kortmann eds. *Cause-Condition-Concession-Contrast: Cognitive and Discourse Perspective*. Berlin: Mouton de Gruyter, 411–437.

Biber, Douglas. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Brown, Penelope and Steven C. Levinson. 1978. Universals in language usage: Politeness phenomena. In Esther N. Goody ed. *Questions and Politeness*. Cambridge: Cambridge University Press, 56–289.

Brown, Penelope and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Brysbaert, Jorina and Karen Lahousse. 2019. Computer-mediated versus non-computer-mediated corpora of informal French: Differences in politeness and intensification in the expression of contrast by au contraire. In Julien Longhi and Claudia Marinica eds. *Proceedings of the 7th Conference on CMC and Social Media Corpora for the*

Humanities. Cergy-Pontoise: The Institute Of Digital Humanities of Cergy-Pontoise University, 48–52.

Cal-Varela, Mario and Francisco Javier Fernández-Polo. 2019. Preparing the ground for critical feedback in online discussions: A look at mitigation strategies. In Julien Longhi and Claudia Marinica eds. *CMC Corpora through the Prism of Digital Humanities*. Paris: L'Harmattan, 15–34.

Cal-Varela, Mario and Francisco Javier Fernández-Polo. 2020. SUNCODAC: A corpus of online forums in higher education. *Nexus-AEDEAN* 2: 44–52.

Cameron, Deborah. 1992. 'Not gender difference but the difference gender makes': Explanation in research on sex and language. *International Journal of the Sociology of Language* 1992/92: 13-26.

Carlo, Jessica Luo and Youngjin Yoo. 2007. "How may I help you?" Politeness in computer-mediated and face-to-face library reference transactions. *Information and Organization* 17/4: 193–231.

Coates, Jennifer. 2004. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language*. New York: Routledge.

Couper-Kuhlen, Elizabeth and Sandra Thompson. 2000. Concessive patterns in conversation. In Elizabeth Couper-Kuhlen and Bernd Kortmann eds. Cause, Concession, *Contrast: Cognitive and Discourse Perspectives*. Berlin: Mouton de Gruyter, 381–410.

Crawford, Mary. 1995. *Talking Difference: On Gender and Language*. London: Sage.

Deris, Farhana Diana, Rachel Tan Hooi Koon and Abdul Rahim Salam. 2015. Virtual Communities in an Online English language learning forum. *International Education Studies* 8/12: 79–87.

Doval-Suárez, Susana M. and Elsa González-Álvarez. 2021. *A Comparison of NS and NNS use of Concessive Markers in Computer- and Non-computer Mediated Discourse*. Paper presented at the *9th International Contrastive Linguistics Conference*, University of Genoa (Italy), 20–21 May 2021.

Fernández-Polo, Francisco Javier and Mario Cal-Varela. 2017. A description of asynchronous online discussions in higher education. In Chelo Vargas-Sierra ed. *Professional and Academic Discourse: An Interdisciplinary Perspective. Epic Series in Language and Linguistics* 2. Alicante: AESLA, 256–264.

Fernández-Polo, Francisco Javier and Mario Cal-Varela. 2018. A structural analysis of student online forum discussions. In Francisco Javier Díaz Pérez and María Águeda Moreno Moreno eds. *Looking at the Crossroads: Training, Accreditation and Context of Use*. Jaén: University of Jaén, 189–200.

Goffman, Erving. 1967. *Interaction Ritual: Essays on Face-to-Face Interaction*. Chicago: Aldine.

Gómez-González, María de los Ángeles. 2017. Concession in evaluative discourse: The semantics, pragmatics and discourse effects of 'but' and 'although'. In Ruth Breeze and Inés Olza eds. *Evaluation in Media Discourse: Rhetoric and Intercultural Pragmatics*. Bern and New York: Peter Lang, 47–80.

Graddol, David and Joan Swan. 1989. *Gender voices*. Cambridge: Basil Blackwell.

Grote, Brigitte, Nils Lenke and Manfed Stede. 1997. Ma(r)king concessions in English and German. *Discourse Processes* 24/1: 87–117.

Guiller, Jane and Alan Durndell. 2006. 'I totally agree with you': Gender interactions in educational online discussion groups. *Journal of Computer Assisted Learning* 22: 368–381.

Hall, Kira. 1996. Cyberfeminism. In Susan Herring ed. *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam:

John Benjamins, 147–170.
Harrison, Sandra. 2000. Maintaining the virtual community: Use of politeness strategies in an email discussion group. In Lyn Pemberton and Simon Shurville eds. *Words on the Web*. Exeter: Intellect Ltd., 69–79.
Herring, Susan C. 1994. Politeness in computer culture: Why women thank and men flame. In Mary Bucholtz ed. *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference*. Berkely: Berkeley Women and Language Group, 278–294.
Herring, Susan C. 1996. Posting in a different voice: Gender and ethics in computer-mediated communication. In Charles Ess ed. *Philosophical Perspectives on Computer-Mediated Communication*. Albany: SUNY Press, 115–145.
Herring, Susan C. 2000. Gender differences in CMC: Findings and implications. *Computer Professionals for Social Responsibility Journal* 18/1: 1–9.
Herring, Susan C. 2023. Grammar and electronic communication. In Carol A. Chapelle ed. *The Encyclopedia of Applied Linguistics*. New York: Wiley-Blackwell, 1–9.
Herring, Susan C. and Sharon Stoerger. 2014. Gender and (A)nonimity in computer-mediated communication. In Susan Ehrlich, Miriam Meyerhoff and Janet Holmes eds. *The Handbook of Language, Gender, and Sexuality*. Malden: Wiley-Blackwell, 567–586.
Hinkel, Eli. 2005. Hedging, inflating, and persuading. *Applied Language Learning* 15/1–2: 29–53.
Holmes, Janet. 1982. Expressing doubt and certainty in English. *RELC Journal* 13/2: 9–28.
Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In Won Kim, Ron Kohavi, Johannes Gehrke and William DuMouchel eds. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington: ACM, 168–177.
Hyland, Ken. 1996. Talking to the academy: Forms of hedging in science research articles. *Written Communication* 13/2: 251–281.
Hyland, Ken. 2005. *Metadiscourse.* London: Continuum.
Izutsu, Mitsuku Narita. 2008 Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics* 40/4: 646–675.
Jonsson, Ewa. 2015. *Conversational writing: A Multidimensional Study of Synchronous and Supersynchronous Computer-Mediated Communication*. Frankfurt am Main: Peter Lang.
Jordan, Michelle E., An-Chih Janne Cheng, Diane Schallert, Kwangok Song, SoonAh Lee and Yangjoo Park. 2014. "I guess my question is": What is the co-occurrence of uncertainty and learning in computer-mediated discourse? *International Journal of Computer-Supported Collaborative Learning* 9: 451–475.
Li, Mimi. 2012. Politeness strategies in wiki-mediated communication of EFL collaborative writing tasks. *The International Association for Language Learning Technology Journal* 42/2: 1–26.
Locher; Miriam. 2004. *Power and Politeness in Action: Disagreements in Oral Communication*. New York: Mouton de Gruyter.
Locher, Miriam A. and Richard J. Watts. 2005. Politeness theory and relational work. *Journal of Politeness Research* 1: 9–33.
Locher, Miriam A. and Richard J. Watts. 2008. Relational work and impoliteness: Negotiating norms of linguistic behaviour. In Derek Bousfield and Miriam A. Locher eds. *Impoliteness in Language: Studies on its Interplay with Power in Theory and Practice*. Berlin: Mouton De Gruyter, 77–99.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text – Interdisciplinary Journal for the Study of Discourse* 8/3: 243–281.

Monaco, Leida Maria. 2017. *A Multidimensional Analysis of Late Modern English Scientific Texts from the Coruña Corpus*. A Coruña: Universidade da Coruña dissertation.

Morand, David A. and Rosalie J. Ocker. 2003. Politeness theory and computer-mediated communication: A sociolinguistic approach to analysing relational messages. In *Proceedings of the 36th Hawaii International Conference on System Sciences.* Hawaii: IEEE. https://doi.org/10.1109/HICSS.2003.1173660

Musi, Elena, Debanjan Ghosh and Smaranda Muresan. 2018. ChangeMyView through concessions: Do concessions increase persuasion? *Discourse and Dialogue* 9/1: 1–21.

Park, Jung-ran. 2008. Linguistic politeness and face-work in computer mediated communication: An application of the theoretical framework. *Journal of the American Society for Information Science and Technology* 59/14: 2199–2209.

Puschmann, Cornelius. 2010. Thank you for thinking we could: Use and function of interpersonal pronouns in corporate web logs. In Heidrun Dorgeloh and Anja Wanner eds. *Syntactic Variation and Genre*. Berlin: Mouton De Gruyter, 167–191.

Pyo, Jihoon and Chung Hyun Lee. 2019. The effects of mitigation strategies instruction in peer response to L2 writing through computer-mediated communication at university level. *Multimedia-Assisted Language Learning* 22/4: 103–133.

Savicki, Victor, Dawn Lingenfelter and Merle Kelley. 1996. Gender language style and group composition in Internet discussion groups. *Journal of Computer-Mediated Communication* 2/3, JCMC232. https://doi.org/10.1111/j.1083-6101.1996.tb00191.x

Schallert, Diane L., Yueh-hui Vanessa Chiang, Yangjoo Park, Michelle E. Jordan, Haekyung Lee, An-Chih Janne Cheng and Kwangok Song. 2009. Being polite while fulfilling different discourse functions in online classroom discussions. *Computers and Education* 53/3: 713–725.

Scott, Mike. 2016. *WordSmith Tools Version 7*. Stroud: Lexical Analysis Software.

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

Swanson, Reid, Brian Ecker and Marilyn A. Walker. 2015. Argument mining: Extracting arguments from online dialogue. In Svetlana Stoyancheve, Safiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington and Malihe Alikhani eds. *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue.* Prague: SIGDIAL, 217–226.

Taboada, Maite and María de los Ángeles Gómez-González. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences* 6/1-3: 17–41.

Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In Jaqueline Bordeau, Jim A. Hendler, Roger Nkambou, Ian Horrocks and Ben Y. Zhao eds. *Proceedings of the 25th International Conference on World Wide Web.* Quebec: International World Wide Web Conferences Steering Committee, 613–624.

Tanskanen, Sanna-Kaisa and Johanna Karhukorpi. 2008. Concessive repair and negotiation of affiliation in e-mail discourse. *Journal of Pragmatics* 40/9: 1587–

1600.
Tet Mei, Kirstie Fung, Su-Hie Ting and Kee-Man Chuah. 2023. Detecting female and male language features in Facebook comments by Malaysian millennial users. *Journal of the Southeast Asian Linguistics Society* 16/1: 37–55.
Thomson, Rob and Tamar Murachver. 2001. Predicting gender from electronic discourse. *British Journal of Social Psychology* 40/2: 193–208.
Uzelgun, Mehmet Ali, Dima Mohammed, Marcin Lewinski and Paula Castro. 2015. Managing disagreement through 'yes, but' constructions: An argumentative analysis. *Discourse Studies* 17/4: 467–484.
Van Nguyen, Long. 2010. Computer mediated collaborative learning within a communicative language teaching approach: A sociocultural perspective. *The Asian EFL Journal Quarterly* 12/1*:* 202–233.
Vinagre, Margarita. 2008. Politeness strategies in collaborative e-mail exchanges. *Computers & Education* 50/3: 1022–1036.
Watts, Richard. J. 1992. Linguistic politeness and politic verbal behaviour: Reconsidering claims for universality. In Richard J. Watts, Sachiko Ide and Konrad Ehlich eds. *Politeness in Language: Studies in its History, Theory and Practice*. Berlin: Mouton de Gruyter, 43–69.
Watts, Richard. J. 2003. *Politeness.* Cambridge: Cambridge University Press.
West, Candace and Don H. Zimmerman.1987. Doing gender. *Gender and Society* 1: 125–51.
Yates, Simon. J. 2001. Gender, language and CMC for education. *Learning and Instruction* 11/1: 21–34.

Corresponding author
Susana Doval-Suárez
University of Santiago de Compostela
Department of English and German
Faculty of Philology
Avda. de Castelao s/n
15782 Santiago de Compostela
Spain
E-mail: susanamaria.doval@usc.es

received: November 2023
accepted: July 2024

RiCL Research in Corpus Linguistics

Detecting emerging vocabulary in a large corpus of Italian tweets

Stefania Spina[a] – Paolo Brasolin[b] – Greta H. Franzini[c]
University for Foreigners of Perugia[a] / Italy
Independent researcher[b] / Italy
Institute for Applied Linguistics, Eurac Research, Bozen[c] / Italy

Abstract – This exploratory study investigates lexical change and innovation in contemporary Italian micro-blogging using a corpus of 5.32 million timestamped and geotagged tweets sampled from the 2022 Italian *Twitter* timeline. We develop a new method to identify 720 unattested forms (347 forms and 373 hashtags) as candidate neologisms. Our results show that orthographic variation, univerbation, suffixation, loanwords and portmanteaus are the most common categories of lexical creation in the data analysed, which appears to be driven by creativity, amusement and attention-seeking behaviour rather than a need for new words to define new objects, events or situations.

Keywords – *Twitter*; social media; corpora; Italian; lexical innovation; language change

1. INTRODUCTION[1]

Lexical innovation is a productive mechanism through which languages evolve (Croft 2000; Labov 2001) and adapt to new sociocultural and technological contexts. It is a crucial process for the survival and vitality of languages, as a living language is such when it is able to accommodate the new needs of its community. Lexical innovation is, therefore, integral to the process of language change, affecting all linguistic levels — phonological, morphological, lexical and syntactic— as well as orthographic aspects of languages. Neologisms are the result of the process of lexical innovation and can be defined as new words not belonging to the vocabulary of a language and not yet recorded in dictionaries or formed by adding new meaning to an already existing word. The process of creating new words follows different steps and usually develops from their initial appearance in specific contexts to their spread to wider domains. This process may end with a final institutionalisation of new word forms (Fischer 1998; Kerremans 2015)

[1] The authors wish to thank the anonymous reviewers for their time and valuable feedback, which helped to improve the quality of the paper.

through their inclusion in dictionaries and consolidation in standard use. However, among the vast number of words that are coined in everyday language use, many remain ephemeral, and only a small number of them become new entries in dictionaries and thus part of the vocabulary. This set of emerging lexical forms, which are only occasionally used for short periods of time and do not systematically enter the vocabulary of a language, are nevertheless of linguistic interest for the insight they give into the lexical innovation mechanisms through which languages evolve.

The process of creating new words can be approached from different standpoints: lexicographical applicability, linguistic phenomena involved, and sources used. Firstly, the process of tracing emerging words has direct lexicographical applications in the creation of neologism dictionaries (e.g., Adamo and Della Valle 2003), which collect new words weaved into daily conversation over a certain period of time, officially including them in the vocabulary of a language. Secondly, the linguistic phenomena leading to the creation of new words, be those involving, among others, derivation, composition or semantic shifting, are of great interest in the field of language change, even when emerging forms are sporadic or do not make it into dictionaries. Thirdly, the choice of sources used to trace the process of lexical innovation has great methodological relevance. Traditionally, newspaper texts have commonly been adopted as reliable sources for new word forms and the study of the lexicon of a language (Marello 2020), as they provide the double benefit of being easily available and quantitatively significant (Adamo and Della Valle 2019). Moreover, newspapers are widely circulated and are commonly transmitters of lexical innovation, both for stylistic reasons and the need to refer to new concepts. Held in high regard in contemporary society, newspapers incite the acceptance and spread of new words.

This study works on the hypothesis that social media represents an opportunity to explore (new) words emerging in everyday interaction, for it provides vast amounts of data produced in real time by a large number of speakers. We test this hypothesis for contemporary Italian with an analysis of emerging vocabulary in a sizeable corpus of tweets. Specifically, we propose a methodology geared towards the detection of emerging lexis and identify 347 word forms and 373 hashtags yet unattested in two of the most up-to-date Italian lexical resources, classifying them into 14 categories of lexical creation.

2. PREVIOUS STUDIES

Research on lexical innovation has produced extensive lexicographical works dedicated to neologisms in many languages, including English (e.g., Algeo 1991; Tulloch 1991; Maxwell 2006), French (e.g., Amar 2010; Des Isnards 2014), and Spanish (e.g., Martí Antonín 1998; Alvar Ezquerra 2003; Moliner 2013). Studies on lexical innovation in Italian boast a long tradition, and have led to the production of several dictionaries or collections of new words (e.g., Migliorini 1963; Scotti Morgana 1981; Lurati 1990; Adamo and Della Valle 2003, 2006, 2008; Bencini and Manetti 2005; De Mauro 2006), as well as a substantial body of research (e.g., Lo Duca 1992; Verardi 1995; Adamo and Della Valle 2003, 2017; Marri 2006, 2018; Frenguelli 2008). The relevance of these lexicographic resources lies not only in the fact that they provide a picture of lexical innovation processes as they occur in language, but also in the role they play in the preservation and documentation of those words in a specific time interval.

One of the fundamental issues faced by lexicography in the study of lexical innovation is the distinction between the notions of 'systemic' and 'occasional' forms in vocabulary (Zgusta 1971) or between 'neologisms' and 'nonce words' (Crystal 1997), the latter denoting occasionalisms not adopted into general use. This distinction is central to lexicographic work and should, in fact, make it possible to select words that have been identified as new and eligible for inclusion in general language dictionaries. Furthermore, this distinction concerns all words hanging between acceptance and disappearance, institutionalisation, and fall into oblivion. In this phase of linguistic stasis, emerging words are placed in an "antechamber of vocabulary" (Verardi 1995:28) and are thus unstable. Neologism dictionaries make room for this instability even when the recorded forms prove to be ephemeral.

It follows that the criteria governing the identification and categorisation of emerging forms as potential neologisms are crucial albeit hard to determine. One of the most widely discussed topics in this regard is the classification of the linguistic processes leading to the creation and spread of new words. Traditionally, research on neologisms acknowledges that the means by which languages enrich their vocabulary are essentially five (e.g., Giraud *et al.* 1971; Guilbert 1975; Zolli 1989):

1. Morphological derivation, that is, the formation of new words from pre-existing lexical elements with the addition of affixes. Examples are *autoregalo* 'gift given to oneself', where the prefix *auto-* modifies the noun *regalo* 'gift'; *prosciutteria*

'ham shop', where the suffix *-eria* modifies the noun *prosciutto* 'ham', or *pigiamone*, where the augmentative suffix *-one* modifies the noun *pigiama* 'pajamas'.

2. Morphological compounding, which is the formation of new words from pre-existing separate words combined to form a new compound word. An example is *contapalle* 'fibber', where the verbal form *conta* 'tells' is coupled with the noun *palle* 'lies'.

3. Reduction or orthographic/phonetic adaptation, that is, the formation of new words through the shortening (e.g., acronyms) or the modification of pre-existing forms. Examples are the acronym *rdc* for *reddito di cittadinanza* 'universal basic income', *csx*, a short form for *centrosinistra* 'centre-left', and *tuitt*, an orthographic variation of the form *tweet* reproducing the Italian pronunciation of the English word

4. Contact, which is the acquisition of new words from other languages or dialects ('borrowing') by adapting them to the paradigms of the target language (adapted loanwords) or by preserving them in their original form.[2] Examples from our corpus are *droppare*, the adaptation of the English verb *drop* to the Italian first conjugation in *-are*, and *fallout*, which is used in its original form.

5. Grammatical or semantic shift: the acquisition of new words through a change of grammatical category or the shift in the meaning of pre-existing forms. Examples are *giornalaia* 'newsagent', used to pejoratively connote a *giornalista* 'journalist', and the verb *cuorare* 'heart', an (incorrect) derivation of the noun *cuore* 'heart'.

Another aspect of lexical innovation widely discussed in previous research concerns the sources used to collect candidate neologisms. As previously mentioned, newspapers are commonly acknowledged as reliable sources for new word forms, as well as one of the most influential agents in the acceptance and dissemination of neologisms. In the last few decades, lexicographic projects have been established to track new words emerging in newspapers. One such project is the *Osservatorio Neologico della Lingua Italiana* (ONLI

[2] While we explicitly exclude dialectal forms from our analysis, examples in our corpus of tweets include *poerannoi* 'poor us' (from the Florentine dialect), *fratm*, an abbreviation of 'my brother' (typical of southern Italy) and *giargiana*, which is used in Milan to denote people who are not from Milan.

2012; Adamo and Della Valle 2019), which has released a database now counting 2,986 new words with definition, date of attestation and first retrieved occurrence in the press.

More recently, with the popularisation of other forms of mass communication and conversational participation, research has stressed the benefits of using social media to track new words emerging in everyday conversation (Rodríguez Arrizabalaga 2021; Würschinger 2021; Tarrade *et al.* 2022). Indeed, the natural ebb and flow of conversation fostered by social media brings out vocabulary approximating the immediacy of spoken interaction (Spina 2016, 2019) and lexical creativity from ordinary users as opposed to inventive journalistic discourse (Eisenstein *et al.* 2014).

A number of recent social media-based studies (Grieve *et al.* 2016, 2018; Kershaw *et al.* 2016) have focussed on the initial phase of the lexical innovation process, that located between a word's creation and first use in a specific context, and its spread in different contexts and potential institutionalisation (Fischer 1998; Kerremans 2015). Another advantage of using social media is that it allows researchers to access unprecedented amounts of conversational data (Spina 2019; Laitinen *et al.* 2020), which can provide a reliable quantitative basis for computations of emerging word forms, thus giving a significant boost to the study of language variation and change (Nguyen *et al.* 2016; Hovy *et al.* 2019).

3. The corpus

To explore evolving lexis in contemporary Italian, we sampled and analysed a dataset of timestamped and geotagged tweets from the Italian *Twitter* timeline spanning the entirety of 2022. The dataset contains 5.32 million tweets authored by 153 thousand unique users, totalling 71.5 million tokens (equivalent to 564 million characters).

4. Method

With the exception of manual annotation, our procedure is structured into a reproducible modular data pipeline. Exclusively relying on Open-Source Software, primarily in the

form of widely recognised Phyton packages and GNU tools, our approach ensures transparency and accessibility.[3]

4.1. Corpus creation and preparation

Using *Twitter*'s advanced search query language,[4] we extracted tweets from the 2022 Italian *Twitter* timeline matching the conditions outlined in Table 1. Tweets can contain geolocation data in two distinct forms: 1) a latitude/longitude pair or 2) an association with a place. A place, in this context, refers to an administrative division or a point of interest and is defined by an ID, a country code, a geographical bounding box, and other metadata. Within our corpus, 99.43 per cent of the tweets are associated with a place, only 0.04 per cent have a latitude/longitude pair, and 0.53 per cent have neither. Despite the higher precision of latitude/longitude pairs, we opted to focus exclusively on places, given that they cover the vast majority of tweets and already include the country code necessary to restrict the data to Italy.

Condition	**Explanation**
Lang: it	Written in Italian
Near:italy	Geotagged near Italy
Since: 2022–01–01	On or after 2022/01/01
Until: 2023–01–01	Before 2023/01/01

Table 1: List of *Twitter*'s search query language conditions defining the Italian *Twitter* timeline of 2022

Tweets consist of an ID, a user ID, a timestamp, the complete text, the previously discussed geolocation data, a list of entities and additional metadata. An entity refers to a character range in the full text labelled by a type (such as url, user mention, hashtag, symbol, or media) and other associated metadata.

Firstly, we extracted all full texts into a flat file, intending to load it into the *AntConc* concordancer (Anthony 2022) to facilitate the subsequent manual annotation process. Next, we introduced entity metadata into the full text as delimiter markers to trick the downstream tokenisation process into breaking these richly structured strings correctly;

[3] The documented source code can be accessed at Brasolin (2023). For a detailed description of the computational processing of the linguistic data, see Brasolin *et al.* (2023).
[4] The official documentation of the query language is available at https://github.com/igorbrigadir/twitter-advanced-search/ and the user interface can be accessed at https://www.twitter.com/search-advanced.

for each entity type, we selected distinct pairs from a set of reserved Unicode code points.[5] Figure 1 provides an example of how this procedure was implemented for hashtag entities.

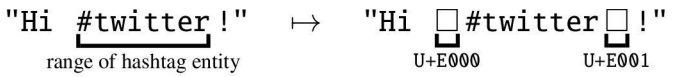

Figure 1: Schematic representation of how we inlined entity range metadata as custom delimiters. This example shows how a hashtag entity is handled

Thirdly, we extracted 5.32 million tweets, preserving their ID, user ID, timestamp, full text with inlined entities, and place ID. Of these tweets, 91.77 per cent are associated with places bearing the IT country code. By aligning their centroids with governmental data,[6] we plotted the tweets containing the emerging forms onto choropleth maps to illustrate the forms' regional distribution across Italy (see Figures 2 and 3 in Appendix A). Specifically, the maps display the simple frequency of each emerging form in the entire corpus (i.e., the sum of the number enclosed in parentheses and, if applicable, that provided in the respective legends) and the number of regional occurrences per million tokens (indicated by the colour scale to the right of the map). Of the remaining tweets, 8.16 per cent are linked to places with other country codes, and 0.07 per cent reference a generic place representative of Italy as a whole. Finally, to tokenise the corpus, we employed the *spaCy v3.6.1* Italian tokeniser.[7]

4.2. Candidate selection

To choose the candidates for annotation we used two different approaches, that is, an already established method in literature and our own attempt at a more interpretable and computationally lighter alternative. This resulted in two groups which have a few candidates in common, as shown in Table 2. The subset of candidates we annotated is the union of the two groups. We now describe both methods in detail.

[5] We picked from the Private Use Area in the Basic Multilingual Plane, which is a set of code points left undefined and reserved for special custom usage (The Unicode Consortium 2022: Chapter 22.5).
[6] Official *ISTAT* data is archived at https://www.istat.it/it/archivio/222527. We used the *GeoJSON* version maintained by the community, available at https://github.com/openpolis/geojson-italy/tree/2023.1.
[7] https://spacy.io/

	Grieve's	Alternate	Overlap	Union
Subset size	6,737	21,132	979	26,890
Fraction of total	0.73%	2.28%	0.11%	2.90%

Table 2: Sizes of the candidate subsets obtained with the two methods, both as a count and as a fraction of the extracted forms. The rightmost columns quantify the size of the overlap and of the union of the two subsets

4.2.1. Grieve's method

The first method is based on previous studies (Grieve *et al.* 2016, 2018) and amounts to calculating how consistently a word's usage increases over time and discarding any word below a certain threshold. The calculation is performed using the Spearman rank correlation coefficient comparing the daily occurrences of a word O (adjusted for the total word count of the day) and the day number. We denote this coefficient ρ_O. The choice for the threshold is somewhat arbitrary. While previous studies, which used much larger datasets, set very high levels at 0.7 and 0.8, we were able to set a lower level due to our smaller dataset and still obtain a manageable number of candidates. We chose $\rho_O > 0.2$, which gave us a subset of 4,090 candidates.

Setting a positive lower limit for ρ_O can penalise usage patterns that could represent an emerging word (for example, a sharp increase in usage before midyear followed by a slow decrease to a stable, non-zero level). Therefore, we decided to include words with $\rho_O < -0.2$ as well, which added 2,336 more potential words to our subset.

In addition, we decided to apply the same calculation to the daily unique users of a word U, obtaining the ρ_U coefficient. We included words with $|\rho_U| > 0.2$, adding 311 more potential words to our subset.

Overall, we selected 6,737 candidates (0.73% of the total) with the following criteria: $max(|\rho_O|, |\rho_U|) > 0.2$.

4.2.2. Alternative method

The measure ρ_O quantifies how much the use of a form increases steadily over the year. As previously discussed, this complex measure aligns with the behaviour of some emerging forms, but it also leaves out possible usage patterns.

We take a different approach and aim to create simple criteria to exclude usage patterns that we would not associate with emerging forms:

a) To rule out accidental and occasional phenomena (like typos, inside jokes, etc.), we set a minimum limit to the count of unique users U and occurrences O.

b) To rule out forms already in use from the past, we set a minimum limit to the day of first occurrence A.

c) To rule out forms that fade away early, we set a high minimum limit to the day of last occurrence Z.

d) To rule out short-lived forms, we set a minimum limit to the length of the usage period $Z - A$.

We chose the following thresholds: $U > 9$, $O > 9$, $A > 7$, $Z > 351$ and $Z - A > 28$. This means we are looking for forms that are used at least ten times by at least ten people, appear from the second week of January, do not disappear before mid December, and last more than four weeks.

The subset defined by the conditions above includes 21,132 candidates (2.28% of the total).

4.3. Corpus annotation

The subset for annotation comprises a total of 26,890 candidates corresponding to 2.90 per cent of the extracted forms. In an effort to streamline the manual annotation process, we used a lexicon of 514 thousand Italian forms (Spina 2014) to automatically filter out attestations from our corpus, resulting in 11,524 candidates.

4.3.1. Non-hashtags

Of the 11,524 candidates, 8133 are non-hashtag forms. The first and second authors of this paper, trained as a corpus linguist and classicist respectively, and manually annotated these forms in two stages. Firstly, we loaded the corpus into AntConc as a flat file and used its *Key Word in Context* tool to look up each form in context. At the same time, we

scanned two freely available online dictionaries, *Garzanti*[8] and *Treccani*,[9] as well as the ONLI neologism database for attestation. The *Slengo* urban dictionary was also consulted for the occasional inspection of slang forms.[10] Based on this comprehensive search, the two annotators categorised forms as either unlikely (assigning them a score of -1) or likely (assigning them a score of 1) to become new dictionary entries, resolving any inter-annotator disagreements through negotiation until consensus was achieved for all forms.

The criteria used to annotate forms as unlikely to become dictionary entries included:

- Attestation in the consulted dictionaries.
- Typos, including those caused by key proximity, e.g., *boungiorno* instead of *buongiorno* 'good morning', *cszzo* instead of *cazzo* 'dick'.
- Established popular neologisms missing from dictionaries, e.g., *bimbominchia* 'sucker'.
- Established foreign words used by the media but missing from dictionaries, e.g., *foliage*, *spending review*, *sponsorship*.[11]
- Nicknames and terms of endearment, e.g., *Gasp* for *Gasperini* or *pupone* 'big baby' for footballer Francesco Totti.
- Vowel elongation for emphasis, e.g., *amooooo* 'loveee'.
- Infrequently used foreign words, e.g., *smoothie*, *veggie*, *waffle*.
- Infrequently used foreign acronyms, e.g., *PTSD*.
- Regionalisms, e.g., *annassero* (Romanesco for *andassero*, third person plural subjunctive of *andare* 'go', *ciolla* 'dick', 'idiot' or 'drugs', depending on the context), *giargiana* (anyone who is not from central Milan)
- Gender-inclusive graphic variants, e.g., *cittadinə* 'citizens'.

In a second stage, we sorted likely candidates according to the ONLI category scheme with minor adjustments and integrations (see Table 3 in Section 5). Specifically, we

[8] https://www.garzantilinguistica.it/
[9] https://www.treccani.it/vocabolario
[10] https://slengo.it/
[11] Gazzardi and Vásquez (2020) provide an overview of studies on the (unnecessary) use of English words in Italian media.

focussed on categories related to formal properties, excluding, for instance, the 'expressive emphasis' category as is commonly found in *Twitter* interactions (Spina 2019) and is inherent in all other categories. Similarly, we merged multiple ONLI categories into one, namely *suffissazione* 'suffixation', *suffissoide* 'suffixoid', *alterazione* 'alteration', *deverbale* 'deverbal' and *denominale* 'denominal' into 'suffixation', and *prefissazione* 'prefixation' and *prefissoide* 'prefixoid' into 'prefixation'. Also, we introduced a new 'tmesis' category to account for forms resulting from the splitting of compounds, e.g., *facenza* from *nullafacenza* 'laziness'. Finally, and where possible, we added the part-of-speech of every form using *TreeTagger's Stein* tagset for Italian as a reference.[12]

4.3.2. Hashtags

Our 11,524 candidates also include 3,391 hashtags. Universally, hashtags appear as either single or unbroken sequences of words (including characters, numerals and underscores), and are often used in their English rendition to expose associated tweets to a wider and more diverse audience.[13] To account for the bias introduced by forced univerbation and English dominance, our hashtag analysis takes a marginally different approach to the one adopted for non-hashtag forms and follows both objective and subjective criteria. We narrow our hashtag selection by filtering out:

1) Those used by nine or fewer distinct users.

2) Proper names, including but not limited to people (e.g., *#gigidagostino*, *#vettel*, as well as portmanteaus like *#basciagoni* used to blend the surnames of Italian *Big Brother* contestants Alessandro Basciano and Sophie Codegoni), places (e.g., *#bozen*, *#regionepuglia*, *#tunisia*), organisations (e.g., *#crocerossaitaliana*, *#aeronauticamilitare*), brands (e.g., *#gucci*, *#versace*), sports teams (e.g., *#acbellinzona*) and events (e.g., *#atpfinals*), festivities (e.g., *#christmas2022*, *#carnevale22*), videogames (e.g., *#eldenring*), music bands (e.g., *#articolo31*) and concerts (e.g., *#cremoninilive22*), films (e.g., *#dontlookup*), and TV shows (e.g., *#1899netflix*, *#cepostaperte*).

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt
[13] See, for example, Hashtagify at https://hashtagify.me

3) Hashtags containing proper names, e.g., *#adaniout* (referring to football commentator Daniele Adani), *#iovotoitaliaviva* 'I support/vote for Italia Viva'.

4) (Combinations of) years, days of the week, times and numbers, e.g., *#8marzo*, *#dicembre2022*, *#anni90*, *#sundaymorning*.

5) Short-lived hashtags relating to a specific incident or time interval, e.g., *#djokovid*, *#draghistan* (referring to former prime minister Mario Draghi's leadership).

6) Univerbated hashtags that we believe have little to no probability of making it into lexical resources, e.g., *#womanlifefreedom*, *#buongiornoatutti* 'good morning everyone'.

We then separate the remaining hashtags into single and univerbated words for manual annotation. The annotation of single-word hashtags, such as *#carobenzina* 'increase in the price of petrol' or *#spiaze* 'it's a pity', is identical to that of non-hashtag forms (see Section 4.3.1), with an additional distinction between *informative* and *evaluative* hashtag function (see Section 6.1) for purposes of analysis. Instead, our annotation of univerbated hashtags, such as *#andratuttobene* 'everything will be alright' or *#booklover*, is objective with respect to ONLI and function categorisations (we do not tag for part-of-speech), but less so in regard to likelihood. In other words, we only consider those univerbated hashtags that we intuitively believe are more likely to establish themselves as new (non-hashtag) forms in Italian social media communication and/or to be acknowledged in authoritative lexical resources, for instance, *#avantitutta* 'let's go!' or *#oldschool*.

5. Results

The selection method described yields a list of 720 emerging forms (347 non-hashtags and 373 hashtags), distributed across 14 categories of lexical creation, as shown in Tables 3 and 4. The emerging forms were also labelled with zero or more part-of-speech tags, producing the distribution shown in Table 5.

The complete list is available in Appendix B, and a machine-readable dataset of the annotated candidates is freely accessible in Franzini *et al.* (2023).

ONLI category	Count	Examples
Orthographic variation	**111**	*Minkiate*, *scienzah*
Suffixation	**60**	*Cinesata*, *adorissimo*
Univerbation	**48**	*Stemmerde*, *massì*
Loanword	**39**	*Reminder*, *scammer*
Portmanteau	**33**	*Lettamaio*, *assurdistan*
Loanword adaptation	24	*Flexo*, *droppare*
Prefixation	8	*Appecoronato*, *iposcolarizzati*
Transcategorisation	7	*Cuora*, *panchinato*
Acronym	6	*Lmv* (li mortacci vostri), *vfc* (vaffanculo)
Compounding	4	*Contapalle*, *cessodestra*
Deonymic derivation	3	*Drum*, *cippalippa*
Redefinition	2	*Maranza*, *giornalaia*
Acronymic derivation	1	*Effeci*
Tmesis	1	*Facenza*
Total non-hashtag forms	**347**	

Table 3: Counts of forms by category, with examples

ONLI category	Count	Examples
Loanword	279	*#aperitif*
Univerbation	50	*#accaddeoggi*
Portmanteau	21	*#caturday*
Acronym	13	*#pdr* (Presidenza della Repubblica)
Compounding	5	*#caroenergia*
Orthographic variation	4	*#povery*
Prefixation	1	*#extraprofitti*
Total hashtag forms	**373**	

Table 4: Counts of hashtag forms by category, with examples

Part of Speech	Non-hashtag	Hashtag	Total
NOM (noun)	201	189	390
ADJ (adjective)	72	23	95
INT (interjection)	46	5	51
VER (verb)	30	17	47
ADV (adverb)	13	1	14
PRO (pronoun)	8	0	8
CON (conjunction)	7	0	7
NPR (name)	5	0	5
PRE (preposition)	2	0	2

Table 5: Counts of PoS tags by form type, and total. Note that forms can have zero or multiple tags

6. Discussion

In the following, we focus on non-hashtag emerging forms, discussing the role of hashtags in a separate section.

The results of the extraction and filtering of emerging forms in the *Twitter* timeline of 2022 allowed us to identify some noteworthy patterns in the mechanisms underlying lexical innovation in Italian. 60 new words (17% of the total number) are formed through suffixation, which is traditionally one of the most common mechanisms languages rely on to create new words (Iacobini and Thornton 1992). These 60 emerging lexical items are mainly created using the derivative suffixes that Italian resorts to in its morphological processes. Examples are the suffixes *-mento* (*impiattamento*, 'plate up'), *-ismo* (*cialtronismo*, behaviour characteristic of a slacker), *-ista* (*abilista*, 'ableist'), *-ato* (*quarantenato*, 'quaranteened'), *-ata* (*poverata*, action characteristic of a poor person), *-eria* (*prosciutteria*, 'ham shop'), *-iolo* (*legaiolo*, hostile designation of a follower of the Italian right-wing populist political party Lega), *-one* (*cazzarone*, 'big/master bullshitter'), and *-azzo* (*coglionazzo*, 'big idiot') or *-ero* (*tuitteri*, '*Twitter* users').

To create new lexical items in Italian, therefore, *Twitter* users rely on established mechanisms. Some, such as derivation through the suffixes mentioned above, are rooted in the earliest stages of the history of the Italian language, whereas others seem to emerge specifically in *Twitter* interactions. An example is the superlative suffix *-issimo*, which is very common in Italian and has the function of intensifying adjectives (Micheli 2020), as in *bellissimo* 'very beautiful'. The suffix *-issimo* has already widened its range of applications, as it can also be found applied to nouns (see Grandi (2017); e.g., *partitissima* 'very important match'.

In our corpus, this suffix finds additional applications. In two of the three emerging forms ending in *issimo* (*adorissimo* and *riderissimo*, see example (1)), the intensifying suffix does not modify an adjective but a verb (*adorare* 'adore' and *ridere* 'laugh'). These two forms represent a further extension of the possible combinations of the suffix *-issimo* and are of major interest because they not only involve lexical but also morphological innovation.

(1) *Io lo* ***adorissimo****, un genio assoluto di simpatia*.
'I adore him so much, an absolute genius in likeability'.

The third new form in *-issimo* detected in our corpus is *incantevolissimissima* 'very very enchanting'. In this case, the form is anomalous for semantic reasons because *-issimo* is applied to an inherently intensified and not gradable adjective.

The search for intensification (Spina 2019) and language economy seems to drive participants in *Twitter* interactions to create new lexical forms. Other examples are instances of the suffix *-one*, for example *cazzarone* 'big/master bullshitter', *rosiconi* (people who feel anger and/or jealousy for someone else's success), *garone* 'big competition' and *fattoni* 'unreliable individuals', 'junkies'. The shift from the original augmentative meaning of *-one* (e.g., *librone* 'big book') to the intensifying, evaluative and pejorative meaning of our examples can be explained through the extension of the suffix's core meaning 'big' to the new meaning of 'intense' (Grandi 2017), or even 'bad'. While this mechanism is not new in Italian derivational morphology, it seems to be one of the most productive ones, partly because the suffix *-one* can be applied to nouns (*garone*) as well as verbs (*rosicone* from *rosicare* 'feel envy').

Another productive suffix for lexical innovation in *Twitter* is *-ata*, which is "one of the most semantically fragmented Italian suffixes" (Grossmann and Rainer 2004: 253). Among the emerging forms in *-ata*, with the exception of those classified as adapted loanwords such as *cringiata* (something embarassing) or *blastata* 'humiliation', 'derision', four cover at least two of the multiple senses of the suffix: in *cinesata*/*cinesate* (to indicate Chinese products), *mandrakata* 'ingenious find', or 'scam' and *poverata* (to denote an action characteristic of a poor person) *-ata* is attached to a nominal animate subject (a Chinese product, Mandrake, a poor person) to connote an action and a negative/pejorative meaning. Example (2) shows this of *cinesata*.

(2) *Beh l'originale è sempre meglio della* ***cinesata****, si sa.*
'Well, everybody knows that the original is always better than the Chinese version'.

The borrowing of foreign words, whether adapted to Italian morphology or not, is another driver of lexical innovation, covering 18 per cent of all of the new forms. The 63 loanwords come from English, with the only exception of *selca* (see example 3), which is a Korean word for *selfie* (*self* + *camera*), and of *matcha*, used to indicate a variety of Chinese green tea or, as the adaptation of the English 'match' to the Italian third person of the present indicative.

(3) *Se non posta un **selca** con i capelli mossi faccio la pazza.*
'If (s)he doesn't post a selfie with wavy hair I'll act crazy'.

English forms imported into Italian can belong to specific lexical domains, such as music (*e.g., djset*, *soundbar*, *soundcheck*) and online environments (e.g., *admin*, *reel*, *twitstar*, *twitterino* '*Twitter* user', *trollino* 'little troll', *trollazzo* 'big nasty troll'), or be part of general everyday use (e.g., *fail*, *flu*, *reminder*, *shoutout*). The abundance of these commonly used words is a notable advantage of using social media conversations among large and diverse groups of ordinary users as a source for lexical innovation. Indeed, while newspapers do contain features of informal everyday speech (Pulcini *et al.* 2012; Marello 2020), articles penned by a limited number of journalists typically employ a more formal vocabulary associated with politics, news reporting or foreign affairs, often detached from everyday use.

One of the differences between direct and adapted loanwords relates to grammatical categories. With the exception of two interjections (*bollox* and *burp*), the former are mainly nouns and adjectives, whereas adapted loanwords —excluding the few nouns adapted through the alterating suffixes *-ino* (*trollini*) or *-azzo* (*trollazzo*), or through the productive suffix *-ata* (*blastata*, *cringiata*), are mainly verbs (*switchare*, *stalkero*, *ghosta*, *flexo*, *droppare*)— conjugated in the first conjugation in *-are*, as is the case for *ghosta* in example (4):

(4) *Ho perso una persona così immatura che **ghosta** invece di dire che non vuole sentirmi più.*
'I have lost a person so immature they'd rather ghost me than say they no longer want to speak to me'.

This difference lies in the fact that the Italian verbal morphology is much more articulated than its nominal morphology, so a verb borrowed from another language must necessarily undergo adaptations in order to become part of the Italian vocabulary. However, in other collections of Italian neologisms based on newspaper articles, such as the ONLI, loanword adaptation does not even exist as a category. Again, adaptations of foreign words to Italian morphology are familiar in register, and thus not suitable to the more formal journalistic style. Two interesting examples of a noun deriving from an adapted loanword are *cringiata* (5) and *blastata* (6), where *cringe* and *blast* become nouns through the addition of the suffix *-ata*.

(5) *La casa di carta coreana la **cringiata** del secolo ora mi dovete spiegare perché.*
'The Korean house of cards is so cringy now you have to tell me why'.

(6) *Mamma mia che **blastata**, me la sto davvero facendo sotto.*
'My goodness what an attack, I was truly scared'.

32 per cent of all of the detected emerging forms were labelled as orthographic variation, which is the most productive category of lexical creation in our corpus. Related research (Grieve *et al.* 2016:110) reports that:

> spelling variation is not generally considered a standard word formation process, as it is not an option in spoken language. From an orthographic perspective, however, these are new linguistic forms.

While our 111 lexical forms mostly align with this observation and are treated as candidates for dictionary inclusion, there are exceptions. Some of their functions are closely tied to the peculiar context of social media interactions, including the need to write quickly and within limited character counts, which often leads to word shortening (e.g., *rix* for *risposta* 'answer'; *sll* for *sullo/a* 'on'; *snx* for *sinistra* 'left'; *csx* for *centrosinistra* 'centre-left'). Similarly, in an effort to conceal potentially offensive or sensitive words, online users often resort to leetspeak to trick automatic censoring filters without altering the words' readability (e.g., *f4scist4* for *fascista* 'fascist', or *merd@* and *merxa* for *merda* 'shit'). However, there are cases of forms labelled as orthographic variation that serve other functions and reveal some interesting driving mechanisms for the creation of new words. An example is orthographic variation used as a joke (e.g., *gomblotto* for *complotto* 'conspiracy', *graduidamende* for *gratuitamente* 'free', *kultura* and *kompagni* for *cultura* 'culture' and *compagni* 'companions/comrades'), or for emphasis (e.g., *coolo* 'arse', *minkiate* 'bullshit (talk/things)', *pikkolo* 'small', *pazzeska* 'crazy'). In all of these cases, the replacement of one or more characters is capable of conveying nuances of meaning that the original spelling could not convey. In *gomblotto*, for instance, the initial *g* alludes to a regional pronunciation of the word; in *kompagni* and *kultura* the letter *k* replaces the *c* to allude to German spelling, and thus to the country's stereotypical authoritarian regime. Moreover, as voth *gomblotto* and *graduidamende* mimic the mispronunciation in spoken Italian of the correct form (be that out of ignorance or dialectal influence), their use moves beyond the confines of written language.

While on the subject of mispronunciation, orthographic variation is also used to mock the Italian pronunciation of foreign words, such as *biutiful* 'beautiful', *singol* 'single', *vairus* 'virus' or *vaucher* 'voucher', and, in a small number of cases, to convey sarcasm. In example (7), the orthographic variation of *scienza* 'science' with the final *-h* serves as a sarcastic expression of scepticism towards scientific advances.

(7) *Credete ciecamente nella **scienzah** anche contro l'evidenza.*
'You blindly believe in pseudoscience against all evidence'.

Univerbation is another productive category of lexical innovation involving the graphic representation of words. In this category, we include all sequences of two or more forms merged by *Twitter* users into a single word through blank space removal, e.g., *buonagiornata* 'goodday' and *ierisera* 'lastnight'. Univerbation has been integral to the evolution of the Italian language over the centuries, leading to the formation of new lexical items in common use today by joining two existing words together (e.g., *invece* 'instead', from the forms *in* and *vece*). Online conversations make frequent use of univerbated forms, partly for a need to economise on the number of characters, and partly owing to hashtags, which —when consisting of two or more words— are necessarily univerbated forms. However, a number of univerbation occurrences in our *Twitter* corpus serve, once more, as an emphatic device, as is the case for *eddaiii* (from *e dai*, 'come on'), *evvaiiiiii* (from *e vai*, 'go/yes'), *opperbacco* (from *o perbacco*, 'my goodness'), *stemmerde* (from *(que)ste merde*, literally 'these shits' to mean 'these arseholes'). The emphatic forms are often characterised by the syntactic doubling of the initial consonant of the second word (e.g., *massì* from *ma* + *sì* 'but yes of course', where the initial *s-* is duplicated).

Portmanteau words or blends (Micheli 2020) also constitute a category in our list of candidate neologisms. In this case, the emerging form is a word combining two or more existing words, as in *presiniente* from *presidente* and *niente* 'a nobody' (referred to a president), *intertristi* from *interisti* and *tristi* 'sad Inter (football club) supporters', or *nazipass* from *nazi* and *greenpass*. in our corpus, portmanteaus mostly relate to politics and are usually used as ironic wordplay (e.g., *lettamaio*, the fusion of politicians Enrico Letta's and Luigi Di Maio's surnames resembling the word *letamaio* 'pigsty'). Additionally, they differ from candidates categorised as compounds: while portmanteaus combine forms where at least one is part of a word (*presi* for *presidente*), compounds result from the juxtaposition of two full words, as is the case of *contapalle* 'fibber' in example (8):

(8) *Grazie è 1 pagliaccio infame **contapalle**, per quello fa ridere.*
'Thanks he's a hateful fibbing clown, that's why he's funny'.

Our list of new forms only includes four compounds (e.g., *fotocazzo* 'dick pic'). This is consistent with the general spread of compounds in Italian, which tends to favour

derivational rather than compositional morphological processes in the formation of new words (Micheli 2020). The ONLI, for instance, includes 430 neologisms obtained through compounding but more than 1,500 obtained through derivation.

Another category used to classify emerging forms is prefixation. Some of the words included in this category are parasynthetic, that is, they involve the addition of both a prefix and a suffix (Micheli 2020), e,g., *appecorato* from *ad-* + *pecora* 'sheep' + *-ato*, used to denote a servile person. We labelled these forms as 'prefixation' since the prefix semantically trumps the suffix (as is also the case for *iposcolarizzati* 'undereducated', where the *ipo-* prefix connotes the low level of education). *Autoregalo* in (9) is one of the eight forms in this category.

(9) *Beh un **autoregalo** per tirarmi un po' su il morale.*
'Well, a self-gift to cheer me up a little'.

In addition to being less common, prefixed forms are not as informal and are less tied to emphasis or irony: the words *biolaboratori* 'biolaboratories', *iposcolarizzati* and *pregirata* 'prerecorded', for instance, pertain to health, education and videomaking respectively, their prefixes used to form domain-specific lexical items rather than wordplay.

The seven forms labelled as 'transcategorisation' (*cuora*, *cuorare*, *cuoro*, *issima*, *issimo*, *panchinato* and *vaffanculi*) relate to three lemmas (*cuorare* 'heart', *panchinare* 'bench', and *vaffanculo* 'fuck you') and to the superlative suffix *-issimo*, used here as an actual word. The verb *cuorare* in example (10) is derived from the noun *cuore* 'heart' to mean 'like' or 'love' and is thus strictlly used in online conversation.

(10) *Non ti **cuoro**, perché non sono d'accordo.*
'I won't heart you because I don't agree'.

In line with the propensity of *Twitter* interactions to use emphatic and intensified forms, *issimo* in its word form occurrence can both strengthen a preceding superlative, as in example (11), or intensify a preceding adjective, as in example (12).

(11) *Ma come fa ad essere bellissimo **issimo issimo** pure vestito da Aladdin?*
'How can he be so so handsome even when dressed as Aladdin?'

(12) *Il prototipo della sinistra intelligente.... direi anche **issima**.*
'The prototype of the intelligent left.... extremely [intelligent], I would add'.

6.1. The role of hashtags

Hashtags are a form of social tagging that allows social media users to incorporate metadata in their posts (Zappavigna 2015). As such, hashtags are able to convey a range of meanings, for they are part of the linguistic structure of online texts whilst providing additional information about them. Owing to this aggregating role, the present study treats these 'super words' as a separate set of emerging lexical items. The total number of hashtags extracted from our *Twitter* corpus as emerging forms is 373, 75 per cent of which are loanwords (single and univerbated). As well as grouping them into the ONLI categories, we tagged hashtags according to their particular function. This has been described by Spina (2019) as either informative if they serve as topic-marker devices (e.g., *#spuntablu* 'blue tick' in example (13)), or as interpersonal/evaluative if they convey the subjective stance of the author (e.g., *#facciamorete* 'together' in example (14)).

(13) *Trovo incomprensibile la polemica per gli 8$ chiesti in cambio della* ***#spuntablu***.
'I really don't understand the controversy surrounding the $8 charge for a #bluetick'.

(14) *Lo diciamo da liberi e pensanti cittadini attivi!* ***#facciamorete****: tutti a votare, senza disperdere voti!*
'We say this as free and rational active citizens! #together: let's all go out and vote without wasting votes!'

The majority of emerging hashtags has an informative function (63%). They are mostly single (*#christmas*, *#olympics*) or univerbated English words (*#weddingday*, *#photooftheday*) used to tag topics. A few widespread acronyms can also be spotted, both from English (*#ootd* for *outfit of the day*) and Italian words (*#rdc* for *reddito di cittadinanza* 'universal basic income'). The rare informative one-word hashtags are compounds, built with the two productive forms *caro-* (*#carobenzina* 'increase in the cost of petrol', *#carobollette* 'increase in household bills' and *#caroenergia* 'increase in energy costs'), and *toto-* (*#totoministri* 'minister pools'). Among the informative and univerbated hashtags based on Italian words, *#allertameteo* 'weatherwarning', *#pausapranzo* 'lunchbreak', and *#biancoenero* 'blackandwhite' are particularly interesting, since they are not restricted to the social media sphere but are used in much more general contexts. Evaluative hashtags are those added to the tweet to comment on its content. They are therefore more creative, starting with their spelling. While we found no instances of orthographic variation in informative hashtags, for evaluative hashtags we

count four, all mostly conveying nuanced ironic meaning. Examples are *#spiaze* 'pity' (15) and *#povery* 'poor people' (16). The former literally means 'feel sorry', but it is used to ironically comment on an unpleasant situation; the original *-c* (*spiace*) becomes *-z* (*spiaze*) to graphically represent the northern pronunciation of a well-known Italian football celebrity from whom the irony originated.

(15) *SerieA: il Cagliari con 1 solo tiro in porta voleva vincerla;* ***#Spiaze****.*
'SerieA: Cagliari wanted to win it with 1 single goal kick; #Pity'.

Similarly, *#povery* (orthographic variation of *#poveri*) adds a touch of British snobbery to the meaning of 'poor':

(16) *Da quello che vedo è più ricco Zhang di voi* ***#povery****.*
'From what I can tell Zhang is richer than you #poorpeople'.

Among the univerbated evaluative hashtags, a number of forms emerge as exhortations (*#andratuttobene* 'everythingwillbealright'), greetings (*#buonagiornata* 'goodday') and interjections (*#buonavita* '[have a] goodlife').

6.2. Institutionalisation in Zingarelli

23 out of the total 347 emerging forms used in *Twitter* in 2022 have been included in *Lo Zingarelli 2024* (Zingarelli 2023), the monolingual dictionary of Italian published in 2023, which incorporates 250 new words and 750 new multi-word forms compared to the previous year's edition. *Lo Zingarelli 2024* can be considered the most up-to-date lexicographical collection of neologisms, partly because the dictionary releases a new edition every year with a section specifically dedicated to neologisms. The 22 forms shared between our candidate neologisms and the last edition of the dictionary (listed in Table 6, below the mid rule) are therefore those that have completed their process of neologisation, from their initial occasional appearance in specific contexts to their spreading in wider situations and, finally, their institutionalisation.

The institutionalised neologisms in our corpus are created through suffixation (12), adapted (4) and direct borrowing (4), prefixation (1), transcategorisation (1), and blending (1). No emerging form created through changes in spelling is accepted into the dictionary the year after its recurring appearance in *Twitter* conversations. This might suggest that orthographic variation is not regarded as a lexicographic criterion which is strong enough for institutionalisation, although the variability in their graphic form is the most common

source of lexical innovation in social media interactions. From a grammatical point of view, the majority of these forms are nouns (12), nouns and adjectives (4), verbs (4) and adjectives (3). It follows that, in the context of *Twitter*, a noun obtained through suffixation seems to be the most likely candidate for dictionary inclusion and, thus, institutionalisation.

Candidate	**Category**	**PoS**
Abilista	Sufixation	ADJ; NOM
Appecoronati	Prefixation	ADJ
Blastata	Loanword adaptation	NOM
Coglionazzo	Sufixation	ADJ; NOM
Condizionalità	Loanword adaptation	NOM
Docuserie	Portmanteau	NOM
Fail	Loanword	NOM
Fallout	Loanword	NOM
Falsona	Sufixation	ADJ; NOM
Fisicati	Sufixation	ADJ
Misunderstanding	Loanword	NOM
Paccare	Sufixation	VER
Paccotto	Sufixation	NOM
Panchinato	Transcategorisation	VER
Pigiamone	Sufixation	NOM
Pigiamoni	Sufixation	NOM
Pirlotto	Sufixation	ADJ
Posturologo	Sufixation	NOM
Rosiconi	Sufixation	ADJ; NOM
Soggettone	Suffixation	NOM
Soundbar	Loanword	NOM
Stalkero	Loanword adaptation	VER
Switchare	Loanword adaptation	VER
#breaking	Loanword	ADJ
#breakingnews	Loanword	N/A
#carobenzina	Compounding	NOM
#crossfit	Loanword	NOM
#genderfluid	Loanword	ADJ
#graphicdesign	Loanword	N/A
#greenwashing	Loanword	NOM
#mindfulness	Loanword	NOM
#omg	Acronym	INT
#reel	Loanword	NOM
#reels	Loanword	NOM
#street	Loanword	NOM
#totoministri	Compounding	NOM

Table 6: Hashtags and non-hashtag forms acknowledged in *Lo Zingarelli 2024* with their respective ONLI category of lexical creation and part(s)-of-speech

7. CONCLUSION

This exploratory study represents the most extensive investigation into lexical innovation in Italian *Twitter* yet. Our findings show that the emergence of new words in *Twitter* appears to be driven more by creativity, entertainment, and a desire for attention rather

than a necessity to introduce novel terms to describe new objects or events. Indeed, the 347 emerging forms mainly perform functions related to irony (*povery*, *presiniente*), intensification (*adorissimo*) and emphasis (*massi*). As has been consistently highlighted in previous studies on social media discourse (e.g., Zappavigna 2012; Spina 2019), the sense of belonging to a large (online) community significantly influences the generation and spread of new words. Some of these coined expressions have the potential of being adopted and reused not only in spoken discourse but also in online communication streams and, in a trans-medial perspective, by the media. The dynamics of their diffusion and a deeper investigation into their probability of becoming institutionalised neologisms could be the focus of future research.

The one-year time frame we adopted proves effective for the detection of emerging usage patterns in the dynamic context of *Twitter*, where linguistic phenomena surface and disseminate rapidly, supporting us in our goal to explore the initial emergence of (novel) words. Nonetheless, it may not capture forms that spread more slowly, maintaining a consistent but slower rate of propagation.

Follow-up work will extend the analysis to additional timelines but, owing to the lately takeover of *Twitter*, which has significantly undermined its value for academic research, will likely have to be redirected to other openly accessible micro-blogging platforms, such as *BlueSky*,[14] or *YouTube* (comments).[15] Furthermore, we will investigate the geographical distribution of emerging forms and hashtags with the aim of identifying regional patterns of lexical creation across Italy. Finally, we will leverage our annotated data to explore how the outcomes of the two methods adopted differ when adjusting threshold choices, aiming to identify optimal points as practical guidelines for future research.

REFERENCES

Adamo, Giovanni and Valeria Della Valle. 2003. *Neologismi Quotidiani. Un Dizionario a Cavallo del Millennio*. Firenze: Leo S. Olschki.

Adamo, Giovanni and Valeria Della Valle. 2006. *Che Fine Fanno i Neologismi? A Cento Anni dalla Pubblicazione del Dizionario Moderno di Alfredo Panzini*. Firenze: Leo S. Olschki.

[14] https://bsky.app/
[15] https://www.youtube.com

Adamo, Giovanni and Valeria Della Valle. 2008. *Le Parole del Lessico Italiano*. Roma: Carocci.

Adamo, Giovanni and Valeria Della Valle. 2017. *Che Cos'è un Neologismo*. Roma: Carocci.

Adamo, Giovanni and Valeria Della Valle. 2019. *Osservatorio Neologico della Lingua Italiana: Lessico Parole Nuove Dell'italiano*. Roma: ILIESI Digitale.

Algeo, John ed. 1991. *Fifty Years Among the New Words. A Dictionary of Neologisms, 1941–1991*. Cambridge: Cambridge University Press.

Alvar Ezquerra, Manuel. 2003. *Nuevo diccionario de voces de uso actual*. Madrid: Arco Libros.

Amar, Yvan. 2010. *Les Mots de L'actualité*. Paris: Éditions Belin.

Anthony, Laurence. 2022. *AntConc (Version 4.2.0)* [Computer software]. https://www.laurenceanthony.net/software.

Bencini, Aandrea and Beatrice Manetti. 2005. *Le Parole Dell'Italia che Cambia*. Grassina: Le Monnier Università.

Brasolin, Paolo. 2023. *Breviloquia Italica: Data Pipeline* (Version 1.1.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10010427

Brasolin, Paolo, Greta H. Franzini and Stefania Spina. 2023. "Ti blocco perché sei un trollazzo": Lexical innovation in contemporary Italian in a large Twitter corpus. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini and Nicole Novielli eds. *Proceedings of the Ninth Italian Conference on Computational Linguistics.* Venice: CEUR-WS. https://ceur-ws.org/Vol-3596/paper12.pdf

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. Harlow: Pearson Education.

Crystal, David. 1997. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

De Mauro, Tullio. 2006. *Dizionarietto di Parole del Futuro*. Roma: Editori Laterza.

Des Isnards, Alexandre. 2014. *Dictionnaire du nouveau Français*. Paris : Allary Éditions.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9/11: e113114. https://doi.org/10.1371/journal.pone.0113114

Fischer, Roswitha. 1998. *Lexical Change in Present-day English: A Corpus-based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Gunter Narr Verlag.

Franzini, Greta H., Stefania Spina and Paolo Brasolin. 2023. *Breviloquia Italica: Annotations (Version 1.0.1)* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10010528

Frenguelli, Gianluca. 2008. Come si studiano le parole nuove. In Maurizio Dardano and Gianluca Frenguelli eds. *L'Italiano di Oggi. Fenomeni, Problemi, Prospettive*. Roma: Aracne, 99–120.

Gazzardi, Antonella and Camilla Vásquez. 2020. A taxonomic approach to the use of English in the Italian media. *World Englishes* 41: 1–14.

Giraud, Jean, Pierre Pamart and Jean Riverain. 1971. *Les Mots dans le Vent*. Paris : Larousse.

Grandi, Nicola. 2017. Intensification processes in Italian: A survey. In Maria Napoli and Miriam Ravetto eds. *Exploring Intensification: Synchronic, Diachronic and Cross-Linguistic Perspectives*. Amsterdam: John Benjamins, 55–77.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2016. Analyzing lexical emergence in modern American English online. *English Language and Linguistics* 21/1: 99–127.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46/4: 293–319.

Grossmann, Maria and Franz Rainer. 2004. *La Formazione delle Parole in Italiano*. Tübingen: Max Niemeyer Verlag.

Guilbert, Louis. 1975. *La Créativité Lexicale*. Paris: Larousse.

Hovy, Dirk, Afshin Rahimi, Timothy Baldwin and Julian Brooke. 2019. Visualizing regional language variation across Europe on Twitter. In Stanley D. Brunn and Roland Kehrein eds. *Handbook of the Changing World Language Map*. Cham: Springer, 3719–3742.

Iacobini, Claudio and Anna M. Thornton. 1992. Tendenze nella formazione delle parole nell'italiano del ventesimo secolo. In Bruno Moretti, Dario Petrini and Sandro Bianconi eds. *Linee di Tendenza Dell'italiano Contemporaneo. Atti del XXV Congresso Internazionale della Società di Linguistica Italiana*. Roma: Bulzoni, 25–55.

Kerremans, Daphné. 2015. *A Web of New Words*. Bern: Peter Lang.

Kershaw, Daniel, Matthew Rowe and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In Paul N. Bennet ed. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. New York: ACM, 553–562.

Labov, William. 2001. *Principles of Linguistic Change*. Malden: Wiley-Blackwell.

Laitinen, Mikko, Masoud Fatemi and Jonas Lundberg. 2020. Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence* 3. https://doi.org/10.3389/frai.2020.00046

Lo Duca, Maria G. 1992. "Parole nuove," regole e produttività. In Bruno Moretti, Dario Petrini and Sandro Bianconi eds. *Linee di Tendenza Dell'italiano Contemporaneo. Atti del XXV Congresso Internazionale della Società di Linguistica Italiana*. Roma: Bulzoni, 57–81.

Lurati, Ottavio. 1990. *3000 Parole Nuove: La Neologia Negli Anni 1980–1990*. Bologna: Zanichelli.

Marello, Carla. 2020. New words and new forms of linguistic purism in the 21st century: The Italian debate. *International Journal of Lexicography* 33: 168–186.

Marri, Fabio. 2006. Parole nuove, meno nuove, troppo nuove (I). *Lingua Nostra* 57/3–4: 113–122.

Marri, Fabio. 2018. I neologismi dentro e fuori dei repertori recenti. *Quaderns d'Italià* 23: 11–26.

Martí Antonín, María A. 1998. *Diccionario de Neologismos de la Lengua Española*. Barcelona: Larousse.

Maxwell, Kerry. 2006. *From Al desko to Zorbing. New Words for the 21st Century*. London: Macmillan.

Micheli, M. Silvia. 2020. *La Formazione delle Parole. Italiano e altre Lingue*. Roma: Carocci editore.

Migliorini, Bruno. 1963. *Parole Nuove: Appendice di Dodicimila Voci al "Dizionario Moderno" di Alfredo Panzini*. Milano: U. Hoepli.

Moliner, María. 2013. *Neologismos del Español Actual*. Madrid: Gredos.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42/3: 537–593.

Osservatorio Neologico della Lingua Italiana (ONLI). 2012. *Parole Nuove dai Giornal*. https://www.iliesi.cnr.it/ONLI/BD.php.

Pulcini, Virgina, Cristiano Furiassi and Félix Rodríguez González. 2012. The Lexical influence of English on European languages: From words to phraseology. In

Cristiano Furiasi, Virginia Pulcini and Félix Rodríguez González eds. *The Anglicization of European Lexis*. Amsterdam: John Benjamins, 1–24.
Rodríguez Arrizabalaga, Beatriz. 2021. Social networks: A source of lexical innovation and creativity in contemporary peninsular Spanish. *Languages* 6/3: 138. https://doi.org/10.3390/languages6030138
Scotti Morgana, Silvia. 1981. *Le Parole Nuove*. Bologna: Zanichelli.
Spina, Stefania. 2014. Il Perugia Corpus: Una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In Roberto Basili, Alessandro Lenci and Bernardo Magnini eds. *Proceedings of the First Italian Conference on Computational Linguistics*. Pisa: Pisa University Press: 354–359.
Spina, Stefania. 2016. Le conversazioni scritte dei social media: Un'analisi multidimensionale. In Francesca Bianchi and Paola Leone eds. *Linguaggio e Apprendimento Linguistico: Metodi e Strumenti Tecnologici*. Milano: Associazione Italiana di Linguistica Applicata, 83–102.
Spina, Stefania. 2019. *Fiumi di Parole. Discorso e Grammatica delle Conversazioni Scritte in Twitter*. Canterano: Aracne editrice.
Tarrade, Louise, Magué, Jean-Philippe and Jean-Pierre Chevrot. 2022. Detecting and categorising lexical innovations in a corpus of tweets. *Psychology of Language and Communication* 26/1: 313–329.
The Unicode Consortium. 2022. *The Unicode Standard* (Version 15.0.0). Unicode Consortium. https://www.unicode.org/versions/Unicode15.0.0/
Tulloch, Sara. 1991. *The Oxford Dictionary of New Words. A Popular Guide to Words in the News*. Oxford: Oxford University Press.
Verardi, Giuseppe Marco. 1995. *Le Parole Veloci. Neologia e Mass Media Negli Anni 90*. Locarno: Armando Dadò.
Würschinger, Quirin. 2021. Social networks of lexical innovation: Investigating the social dynamics of diffusion of neologisms on Twitter. *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.648583
Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media. How We Use Language to Create Affiliation on the Web*. London: Continuum.
Zappavigna, Michele. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25/3: 274–291.
Zingarelli, Nicola. 2023. *Lo Zingarelli 2024: Vocabolario della Lingua Italiana*. Bologna: Zanichelli.
Zgusta, Ladislav. 1971. *Manual of Lexicography*. The Hague: Mouton De Gruyter.
Zolli, Paolo. 1989. *Come Nascono le Parole Italiane*. Milano: Rizzoli.

Corresponding author
Stefania Spina
University for Foreigners of Perugia
Department of Italian Language, Literature and Art in the World
Piazza Fortebraccio, 4
06123 Perugia
Italy
E-mail: stefania.spina@unistrapg.it

received: November 2023
accepted: July 2024

APPENDIX A: CHOROPLETH MAPS

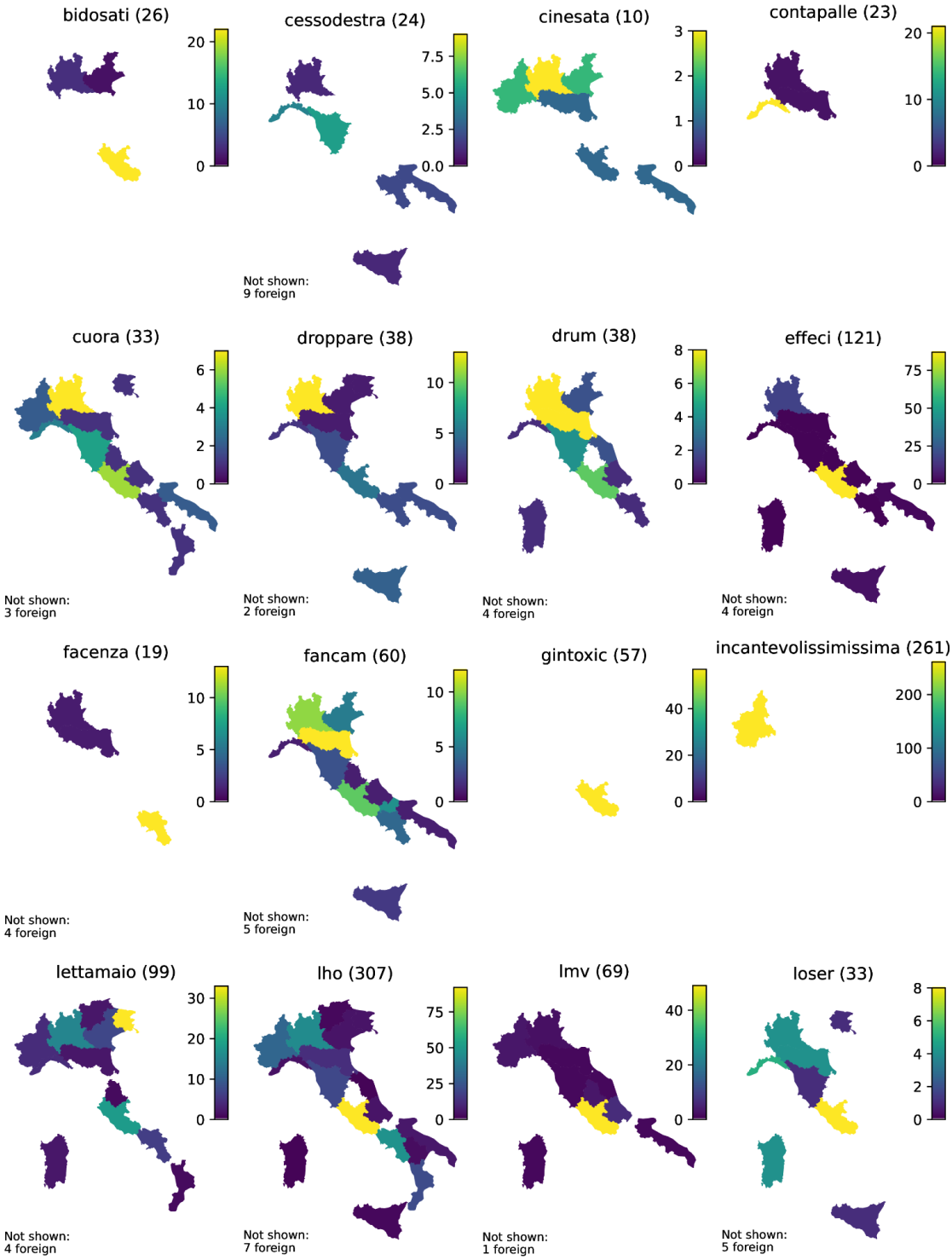

Figure 2: Choropleth maps of candidate neologisms from A to L. The colour scale represents instances per million tokens at the regional level. Total occurrences in Italy are provided with the titles. Occurrences outside Italy are not shown and counted in the legends.

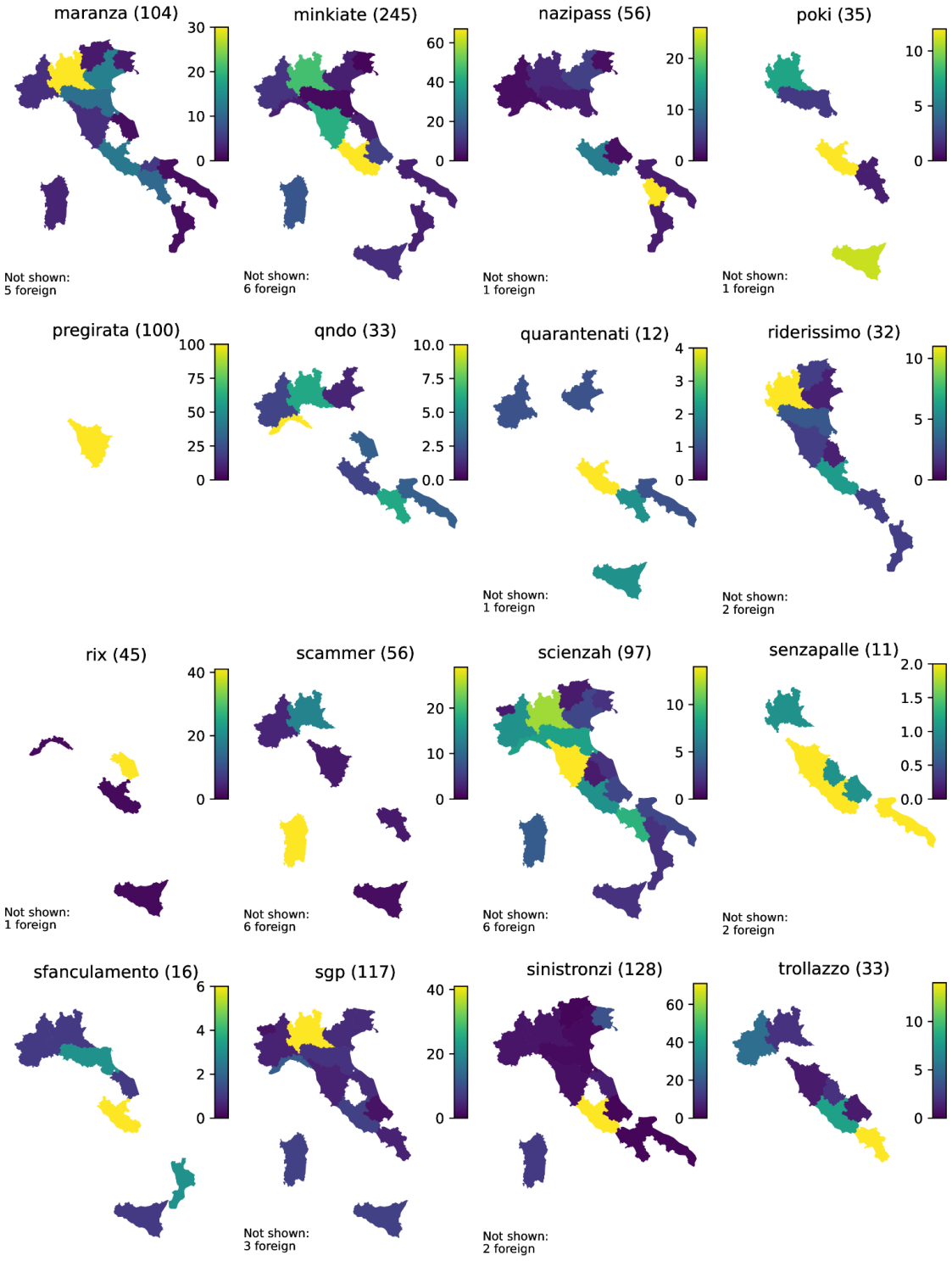

Figure 3: Choropleth maps of selected candidate neologisms from M to Z. The colour scale represents instances per million tokens at the regional level. Total occurrences in Italy are provided with the titles. Occurrences outside Italy are not shown and counted in the legends

APPENDIX B: FULL LIST OF EMERGING FORMS

B.1. Non-hashtag forms by category

Orthographic variation (111): *5s, accaunt, adovo, affan, amerika, amiketti, amio, amïo, ancielo, anzia, assaj, azzzzz, babbà, benza, biutiful, c4zz0, c@@@o, caiser, cazxi, cazza, cme, collab, comple, coolo, csx, cuxo, dll, duddi, eu4ia, f4scist4, f4scista, fassisti, feffettissimo, gaz, gomblotto, graduidamende, graduidamente, graduido, gretina, grin, incaxxano, incaz, incazz, kaffè, kaimano, kazzate, kompagni, kultura, laik, leccac, lvi, madreh, mbeh, mer*a, merd@, merxa, minkiate, minkione, neanke, nerah, norde, nsomma, okk, okok, ovvove, pazzeska, pienah, pikkolo, pk, plis, poki, qlcosa, qlcuno, qlk, qndo, qnt, qt, qulo, qusto, reposta, rimba, rix, rubba, scienzah, sexi, sexo, singol, sinix, sll, snx, stronxate, stronz, tks, troya, trq, tuitt, ubri, urka, vafancul, vaff, vaffan, vaffanc, vairus, vaucher, vergonya, xazzo, xe, xhe, xsino, yessa, zola.*

Suffixation (60): *abilista, accannate, accannato, adorissimo, amorina, baguettari, benissimamente, busoni, cazzarone, cazzaroni, ciacchera, cialtronismo, cinesata, cinesate, coglionazzo, ducessa, eurini, falsona, fattoni, fisicati, garone, godicchio, gretini, impiattamento, incantevolissimissima, legaiolo, mandrakata, memiamo, paccare, paccotto, patati, patatino, personaggione, piagnina, piddini, pigiamone, pigiamoni, pirlotto, pisellate, posturologo, poverata, presidenta, prezzemolina, prosciutteria, quarantenati, riderissimo, ridolini, rosiconi, senzadubbiamente, sfanculamento, sierare, sierata, soggettone, tridosato, triplodosati, tuitteri, twettini, twitteri, zanzarologi, zanzarologo.*

Univerbation (48): *ammiocuggino, anchio, buonagiornata, buonamattina, buontutto, cho, ciaobuogiorno, daltronde, demmè, diobono, dioca, diocan, dioporco, eddaiii, eropd, essu, estigrancazzi, evvaiiiiii, flattax, fuoriluogo, gintonic, graziealcazzo, ierisera, instagramstory, lho, lowcost, massí, masticazzi, mavalà, mavattelapijànd', miocuggino, miraccomando, ncazzo, nculo, noeuro, nowar, opperbacco, porcaputtana, porcodd, senzapalle, serietv, sottocasa, stemmerde, stica, streetart, terzopolo, tuttappost, ziocane.*

Loanword (39): *admin, af, baller, banger, bollox, burp, champ, cishet, dilf, djset, drip, fail, fallout, fanbase, fancam, flu, horny, locals, loser, mentor, misunderstanding, reel, reminder, rimming, scammer, selca, shoutout, showrunner, slim, solution, soundbar, soundcheck, stats, terf, throwback, tier, topping, twitstar, venue.*

Portmanteau (33): *5scemi, 5stalle, assurdistan, deltacron, docuserie, estaters, fasciocomunista, fascioleghista, fascioleghisti, flurona, gintoxic, giornalanza, grillioti, grillopiddini, grillopitechi, intertristi, inverners, lettamaio, nazipass, naziucraini, pdiota, pdioti, piddiota, piddioti, pidiota, pidioti, presiniente, putler, renziota, renzioti, scansuolo, sinistronzi, tecnopolo.*

Loanword adaptation (24): *blastata, blessata, boyz, broder, condizionalità, cringiata, droppare, eppi, flex, flexo, followo, ghosta, matcha, pullato, schip, squirtare, stalkero, switchare, trollata, trollazzo, trolling, trollini, twerka, twitterino.*

Prefixation (8): *appecorato, appecoronati, autoregalo, bidosati, biolaboratori, intrasezioni, iposcolarizzati, pregirata.*

Transcategorisation (7): *cuora, cuorare, cuoro, issima, issimo, panchinato, vaffanculi.*

Acronym (6): *afc, lms, lmv, rdc, sgp, vfc.*

Compounding (4): *cessodestra, contapalle, fotocazzo, fregacazzi.*

Deonymic derivation (3): *cippalippa, drum, lippa*

Redefinition (2): *giornalaia, maranza.*

Acronymic derivation (1): *effeci.*

Tmesis (1): *facenza.*

B.2. Emerging hashtag forms by category

Loanword (279): *#actor, #adoptdontshop, #adventure, #airport, #amazing, #aperitif, #archaeology, #artist, #artistic, #artwork, #attitude, #autumn, #autumnvibes, #award, #awards, #babyboy, #baroque, #beard, #behappy, #bestfriends, #bicycle, #biodiversity, #birds, #black, #blackandwhite, #booklover, #breaking, #breakingnews, #budgetcap, #burger, #butterfly, #cancer, #cathedral, #catlife, #catlover, #chess, #chill, #circulareconomy, #cityscape, #climate, #climateaction, #climatechange, #clubbing, #coffeelover, #colorful, #colour, #colours, #comedy, #communication, #couple, #cousins, #creativity, #crossfit, #cryptocurrency, #culturalheritage, #curvy, #cycling, #dad, #dancers, #daughter, #dawn, #daytime, #devotion, #digitalart, #dinnertime, #documentary, #doglover, #drama, #dress, #dusk, #earth, #earthquake, #ebike, #elegance, #euphoria, #fail, #fairplay, #fall, #familyfirst, #fashionstyle, #finance,*

#followme, #followme (unicode homograph of the previous entry), #foryou, #freetime, #fridayvibes, #fuck, #fuckcancer, #gameday, #genderfluid, #getoutthere, #glasses, #goalkeeper, #goat, #gold, #goodevening, #goodtimes, #graphicdesign, #grateful, #gratitude, #greenwashing, #gymlife, #hair, #hairstyle, #happybday, #happyholidays, #happyness, #hat, #health, #heart, #holiday, #homedecor, #homedesign, #hospitality, #icecream, #ink, #innovation, #instore, #interior, #interiordesign, #interview, #investing, #investment, #iphonography, #italiansdoitbetter, #journalism, #journey, #joy, #kids, #landscapes, #life, #lighting, #lights, #likeforlikes, #lunchtime, #luxury, #macteanimo, #marathon, #medieval, #meditation, #menstyle, #mentalhealth, #midnights, #migrants, #mindfulness, #mirror, #mondaymood, #monochrome, #monument, #musiclover, #naturalbeauty, #naturelovers, #newbook, #newcollection, #newlife, #newlook, #nextgen, #nightlife, #nomask, #noracism, #novax, #nowar, #nowars, #nowplaying, #nowwatching, #oldschool, #olympics, #onelove, #onfire, #partytime, #peaceandlove, #peacenotwar, #philosophy, #photoart, #photograghy, #photographer, #photooftheday, #picoftheday, #pictures, #pizzatime, #pontifex, #portrait, #portraits, #positivevibes, #prayforpeace, #president, #pricecap, #production, #proud, #quality, #quoteoftheday, #quotes, #rain, #raw, #recording, #reel, #reels, #relaxing, #remember, #renaissance, #rescue, #respect, #roadtrip, #roses, #sad, #sand, #saturdayvibes, #savetheplanet, #seafood, #seascape, #see, #shadows, #shame, #ship, #shoes, #shoot, #singer, #sisters, #slavaukraini, #slavaukrainii, #slavaukraïni, #song, #songs, #songwriter, #space, #specialguest, #spring, #springtime, #steak, #stopwar, #street, #summercamp, #sunglasses, #supergreenpass, #tatoo, #tattooart, #theater, #thebadguy, #thoughts, #throwbackthursday, #tourism, #town, #trail, #trailrunning, #travel, #travelgram, #traveller, #travelling, #tree, #trees, #tuesdayvibe, #tuscanygram, #vacation, #vanlife, #vibes, #vintagestyle, #viral, #voice, #volcano, #waiting, #wakeup, #walking, #wall, #wanderlust, #war, #waterfall, #waves, #weather, #webmarketing, #weddingday, #whatelse, #wildlife, #win, #window, #wine, #winetime, #winteriscoming, #woman, #women.

Univerbation (50): *#accaddeoggi, #allertameteo, #amoremio, #andratuttobene, #aperitivotime, #avantitutta, #avantiunaltro, #bellavita, #biancoenero, #buonacena, #buonagiornata, #buonappetito, #buonascuola, #buonaserata, #buonavita, #buonefeste, #buonenotizie, #buonevacanze, #buonlavoro, #buononomastico, #buonpranzo, #casadolcecasa, #cessateilfuoco, #ciaociao, #dallapartegiusta, #dalleparoleaifatti,*

#facciamorete, #governodegliorrori, #governodeimigliori, #governodeipeggiori, #governodellavergogna, #governodipagliacci, #grandebellezza, #grazieatutti, #idearegalo, #iomivaccino, #ionondimentico, #iononmollo, #maimollare, #neiperte, #nonato, #nopos, #oggicosi, #pausapranzo, #perte, #qrcode, #romanzoquirinale, #spuntablu, #sulserio, #unovaleuno.

Portmanteau (21): *#bookstagram, #catstagram, #caturday, #chilhavister, #fantacitorio, #farsopoli, #foodstagram, #instaart, #instabook, #instacat, #instadog, #instagood, #instamoment, #instamood, #instaphoto, #instapic, #instatravel, #lettamaio, #pfizergate, #sapevatelo, #sivax.*

Acronym (13): *#bnw, #fyp, #ia, #ig, #mma, #omg, #ootd, #otnba, #pdr, #rdc, #tb, #tbt, #wwiii.*

Compounding (5): *#carobenzina, #carobollette, #caroenergia, #cinesalvini, #totoministri.*

Orthographic variation (4): *#anala, #chesucc3de, #povery, #spiaze.*

Prefixation (1): *#extraprofitti.*

RiCL Research in Corpus Linguistics

Nonbinary pronouns in *X* (*Twitter)* bios: Gender and identity in online spaces

Lucía Loureiro-Porto – José Luis Ariza-Fernández
University of the Balearic Islands / Spain

Abstract – This study explores the usage of nonbinary pronouns on *X* (formerly known as *Twitter*), focusing on THEY and neopronouns like ZE or XE within the nonbinary community. Building on the increasing practice of sharing pronouns, especially in online spaces, the research collects 1,980 *X* accounts using *Followerwonk*. Despite ideological differences across U.S. regions, no substantial variations in pronoun usage are observed. Notably, a preference for rolling pronouns (e.g., *they/she*) emerges, with fewer instances of monopronoun usage (e.g., *they*). When a single pronoun is chosen, it is often accompanied by the respective accusative form, while rolling pronoun users tend to omit the accusative. Users with binary pronouns often prioritize it as their first chosen pronoun. THEY remains the predominant nonbinary pronoun, with neopronouns being rare. The study highlights *X* profiles as valuable sources for understanding linguistic patterns related to social trends, particularly in the context of gender equality and network relations.

Keywords – nonbinary pronouns; singular THEY; neopronouns; gender-inclusive language; social media; *X (Twitter)*

1. INTRODUCTION[1]

The exploration of pronouns as tools for self- and other-reference has received considerable attention in recent decades, primarily through the lens of feminist inquiry (pioneered by Bodine 1975) and, more recently, queer perspectives (e.g., McLemore 2015; Zimman 2017; Bradley 2020; Konnelly and Cowper 2020). The pronoun THEY initially sparked debate due to its role as a singular gender-neutral pronoun, skillfully sidestepping gender assignment, as seen in examples like *someone lost their keys* (Balhorn 2009; Paterson 2014; LaScotte 2016; Loureiro-Porto 2020). However, its evolution expanded beyond gender neutrality to represent nonbinary identities (Bradley *et al.* 2019; Conrod 2019; Bradley 2020; Hekanaho 2020, 2024).

[1] For financial support Lucía Loureiro-Porto is grateful to the *Spanish Ministry of Science, Innovation and Universities*, grant PID2020-117030GB-I00, funded by MICIU/AEI/10.13039/501100011033. Thanks are also due to two anonymous reviewers and the editors of this special issue, whose comments have improved the original version of this manuscript to a large extent. Needless to say, errors or omissions that remain are our responsibility.

Recent research highlights the discomfort of nonbinary individuals, who diverge from the gender binary, grammatically expressed by HE or SHE, resulting in intentional and unintentional misgendering (Simpson and Dewaele 2019: 105–106; Konnelly *et al.* 2024: 453–454). Responding to this, the groundwork laid by feminists for singular THEY made it the prime candidate to fill this void, leading to its recognition as the word of the year in 2019 by Merriam Webster (Harmon 2019).[2] Simultaneously, new alternatives, termed neopronouns, like ZE and XE, emerged to address this gap (Hegarty *et al.* 2018: 55), as illustrated in (1) and (2):

(1) *Clo loves **zir** mother.* (From Hekanaho 2020: 5)

(2) *Terry was going out but **xe** could not find **xir** keys*. (From Hekanaho 2020: 273)

The plethora of emerging pronominal possibilities underscores the complexity of transforming English into a more inclusive language. Nonbinary individuals, recognizing the pivotal role of pronouns in defining their identities, emphasize the significance of being referred to by pronouns that align with their sense of self. Some scholars, such as Zimman (2017: 156), advocate for an egalitarian approach, proposing that the most inclusive method for personal pronoun reference is to inquire directly about individuals' preferred pronouns. Conversely, some argue that certain LGBTQI+ individuals perceive gender pronouns as limiting in encapsulating their complex identities, leading to a call for the complete avoidance of gender-specific pronouns in reference to any individual (Dembroff and Wodak 2018: 372). These discussions illuminate the identity-building function of pronouns, emphasizing their role in intersubjective identity construction through discourse interaction (Bucholtz and Hall 2010; Hekanaho 2024).

In situations where individuals are not explicitly asked about their pronouns, they may choose to overtly state them, as observed in social networks like *X* (formerly *Twitter*), where users have at their disposal 160 characters to define their public profiles (known as bios), according to their own wishes.[3] A cursory examination of random profiles reveals a diverse array of pronoun claims and combinations, including binary pronouns, nonbinary (NB) pronouns, and a blend of binary and NB pronouns, commonly

[2] Whilst we are writing this paper, the Spanish *Real Academia de la Lengua Española* (RAE 2023) announces that one of the new entries added to its electronic version 23.7 is precisely *no binario* 'nonbinary', which constitutes just another piece of evidence that standardizing institutions acknowledge the need to find specific vocabulary to refer to nonbinary individuals.

[3] Referring to those individuals by the pronouns they go by would then constitute an example of good manners, although the social network *X* has lately witnessed a sort of heated debate regarding this issue (Ingram 2023).

referred to as rolling pronouns (e.g., *they/he*; LGBTQ Nation 2022). Moreover, online spaces like *X* and *Tumblr* have been found to favor the diffusion of new pronouns (King and Crowley 2024: 79–82). These social media have also served as battlegrounds for intense discussions surrounding the ideological implications of adopting NB pronouns, as the act of disclosing one's pronouns has "politicized as belonging to the left in current US politics" (King and Crowley 2024: 82). Against this backdrop, this paper conducts an analysis of NB pronoun usage in *X* bios in US-based accounts, considering various intra- and extra-linguistic features, detailed in Section 3 below, with the overarching goal of answering the following research questions:

RQ1: Which NB pronouns are predominantly used in *X* bios?

RQ2: Do NB pronouns coexist with binary ones, and if so, what is the prevalence of each pronoun?

RQ3: Does the claiming of pronouns allow for inflectional morphology (i.e., are non-nominative forms listed)?

RQ4: Are there discernible differences, considering the ideological value of NB pronouns, between individuals residing in cities with a tradition of Republican governments and those in cities with a tradition of Democrat governments?

RQ5: Does the assertion of NB pronouns correlate with specific profiles, such as activism of any sort?

To achieve these objectives, the following sections of the paper unfold as follows: Section 2 outlines the theoretical background, Section 3 explains the methodology, Section 4 reveals the findings, and Section 5 offers a comprehensive discussion. The paper concludes with key insights and conclusions in Section 6.

2. NONBINARY PRONOUNS IN ENGLISH

For over 150 years, English wordsmiths have attempted to establish a gender-neutral pronoun without success (Baron 2010: n.p.), in contrast with some languages that have recently embraced gender-inclusive language approaches and alternatives to binary pronouns have been established, such as *hen* in Swedish, which reflects a growing acknowledgment of gender diversity (Lindqvist *et al.* 2019). Despite the historical existence of non-conforming gender individuals, who have been marginalized and

persecuted for centuries (Herdt 1996: 11), they have faced a persistent lack of visibility and recognition. This is reflected in language, where the absence of an established third person singular genderless pronoun leads to misgendering (i.e., an erroneous attribution of gender, McLemore 2014: 53; see also Hekanaho 2020: 197) for those who do not conform to the gender binary. In this scenario Sections 2.1 and 2.2 review the pronominal choices available for nonbinary individuals and their relative success in recent years.

2.1. NB THEY

Despite the widespread belief that singular THEY is a modern linguistic innovation, its usage was prevalent in written English even before the twentieth century, with the first recorded instances dating back to Old English (Bodine 1975: 131; Curzan 2003: 70–71; Laitinen 2024: 36–38). However, the proscription against using singular THEY due to a lack of number agreement with the singular antecedent became prominent with the advent of prescriptive usage guides in 1770 (*HUGE-database, Hyper Usage Guide of English*; Straaijer 2014). This prohibition persisted until the twenty-first century, as seen in Batko (2004: 118–122), who cautioned against using "everyone...their" in formal speech or writing, advocating awareness of alternatives that adhere to prescriptive rules.

Amidst this prescriptivist landscape, the feminist movement of the 1960s, particularly second-wave feminism, played a pivotal role in revitalizing the usage of singular THEY. This resurgence aimed to combat linguistic sexism, bringing singular THEY into debate and gaining acceptance for referring to antecedents of unknown or irrelevant gender (Balhorn 2009; Paterson 2011, 2014; LaScotte 2016). Consequently, the trajectory of singular THEY being used with singular antecedents dates back to medieval times, where genderless or unknown antecedents were commonly referred to by singular THEY and combined with HE OR SHE (see Baron 2018, for example). Grammarians of that era criticized both options, deeming the first inaccurate due to a lack of number agreement and the second as "clumsy and pedantic" (Bodine 1975: 170; Paterson 2014: 123).

The prescriptive pressure on the use of singular THEY persisted over time, earning it the moniker of an "old chestnut," frequently cited in usage guides (Tieken-Boon van Ostade 2020: 26; 58 out of 77 guides in the *HUGE-database* mention this issue). Nevertheless, the social rejection of generic HE in the late twentieth century, driven by the recognition that a pronoun cannot be simultaneously masculine and generic, led to a

shift in perception. Singular THEY, along with the combination of HE OR SHE, came to be viewed as gender-inclusive and, consequently, the preferred choice among speakers (LaScotte 2016: 63).

This capacity to denote singular antecedents whose gender is unknown or irrelevant likely facilitated the recent adoption of THEY as a choice for referring to nonbinary individuals. This category encompasses those who may not conform to the gender binary, identify with none or both genders, or reject the notion of having a gender identity (Matsuno and Budge 2017: 116). While resistance persists, possibly due to the blurred lines between grammar and social meaning (Konnelly and Cowper 2020: 16), studies have demonstrated the viability of THEY as a NB pronoun (Parker 2017; Lund Eide 2018; Bradley 2019; Hekanaho 2020; among many others). Notably, nonbinary THEY, encompassing inflectional forms such as *they, them, their, theirs,* and *themself*, has gained official recognition from institutions such as the University of Vermont (Scelfo 2015: n.p.) and is listed as a NB pronoun in the 2019 edition of the *Merriam-Webster Dictionary* (Merriam-Webster 2019). It is essential to acknowledge, however, that THEY is not the exclusive contender for an established NB pronoun, as various alternatives have been proposed, as explored in Section 2.2.

2.2. Neopronouns

In addition to the emerging use of THEY as a NB pronoun, the linguistic landscape has seen the introduction of numerous newly coined pronouns in recent decades, collectively referred to as 'neopronouns'. These innovative pronoun sets, still in the process of gaining widespread acceptance, are cataloged on reference sites like http://www.pronouns.org/. The existence of these neologisms could be considered to challenge the conventional belief that pronouns constitute a closed class (Huddleston and Pullum 2002: 425), and, although their success, unlike that of singular THEY, has been limited (Lund Eide 2018; Parker 2017, cited in Hekanaho 2020: 39; Bradley *et al.* 2019), this has not hindered speakers from engaging in continual linguistic innovation. Consequently, the list of neopronouns is extensive and subject to change over time. While acknowledging the absence of a comprehensive academic list, we present here a compilation of "artificial and proposed epicene pronouns" as found in Wikipedia as of 20 November 2023:

	Firstly attested	Nominative	Accusative	Dependent Genitive	Independent genitive	Reflexive
THON	1884	***thon*** is laughing	I called ***thon***	***thons*** eyes gleam	that is ***thons***	thon likes ***thonself***
E	1890	***e*** is laughing	I called ***em***	***es*** eyes gleam	that is ***es***	e likes ***emself***
AE	1920	***ae*** is laughing	I called ***aer***	***aer*** eyes gleam	that is ***aers***	ae likes ***aerself***
TEY	1971	***tey*** is laughing	I called ***tem***	***ter*** eyes gleam	that is ***ters***	tey likes ***temself***
XE	1973	***xe*** is laughing	I called ***xem/xim***	***xyr/xis*** eyes gleam	that is ***xyrs/xis***	xe likes ***xemself/ximself***
TE	1974	***te*** is laughing	I called ***tir***	***tes*** eyes gleam	that is ***tes***	te likes ***tirself***
EY	1975	***ey*** is laughing	I called ***em***	***eir*** eyes gleam	that is ***eirs***	ey likes ***emself***
PER	1979	***per*** is laughing	I called ***per***	***per*** eyes gleam	that is ***pers***	per likes ***perself***
VE	1980	***ve*** is laughing	I called ***ver***	***vis*** eyes gleam	that is ***vis***	ve likes ***verself***
HU	1982	***hu*** is laughing	I called ***hum***	***hus*** eyes gleam	that is ***hus***	hu likes ***humself***
E	1983	***e*** is laughing	I called ***em***	***eir*** eyes gleam	that is ***eirs***	e likes ***emself***
ZE, MER	1997	***ze*** is laughing	I called ***mer***	***zer*** eyes gleam	that is ***zers***	ze likes ***zemself***
ZE, HIR	1998	***ze*** is laughing	I called ***hir***	***hir*** eyes gleam	that is ***hirs***	ze likes ***hirself***
SIE, HIR	2001	***sie*** is laughing	I called ***hir***	***hir*** eyes gleam	that is ***hirs***	sie likes ***hirself***
SEY, SEIR, SEM	2013	***sey*** is laughing	I called ***sem***	***seir*** eyes gleam	that is ***seirs***	sey likes ***Sem self***
FAE	2020	***fae*** is laughing	I called ***faer***	***faer*** eyes gleam	that is ***faers***	fae likes ***faerself***

Table 1: List of proposed neopronouns (adapted from Wikipedia 2023)[4]

The pronouns listed in Table 1 exhibit varying degrees of popularity, with some receiving more attention on authoritative websites like gendercensus.com (2022). Notably highlighted are the following: (1) E (*e/em/eir/eirs/emself*; known as 'Spivak pronouns');[5] (2) EY (*ey/em/eir/eirs/emself*, known as 'Elverson pronouns');[6] (3) ZE (*ze/hir/hir/hirs/hirself*); (4) XE (*xe/xem/xyr/xyrs/xemself*); and (5) FAE (*fae/faer/faer/faers/faeself*) (gendercensus 2022; see also Venkatraman 2020). These pronouns do not only differ in popularity but also in phonological weight: E and EY contain vocalic sounds resonant with SHE and THEY while XE and ZE are sometimes pronounced as /zi:/ or /ksi:/ (Hekanaho 2020: 4).

[4] In fact, Wikipedia lists some sources for each of the pronouns, but many of them are debatable and, with the aim of keeping the explanation simple, we have decided just to include the first attestation date as currently found in the entry. The Wikipedia list of neopronouns is considerably shorter than that proposed by Baron (2020), which contains over 200 possibilities (Stormbom 2024: 416), as well as other compilations available on online platforms such as *Pronouns.page* (featuring 19 neopronouns) and *Pronouny* (which documents over one thousand neopronouns). Consequently, the 16 neopronouns outlined in Table 1 can be confidently regarded as the most commonly utilized sets of NB pronouns.

[5] The term 'Spivak pronouns' is attributed to the mathematician Michael Spivak, who first used *e/em/eir/eirs/emself* in his book *The Joy of TEX: A Gourmet Guide to Typesetting with the AMS-TEX Macro Package* (see *Pronouns.page* 2024).

[6] This term originates from Christine M. Elverson, who won a contest in 1975 with the intention of offering an alternative to singular THEY (see *Pronouns.page* 2024).

Additionally, FAE stands out as it can be considered a nounself pronoun, a category of new pronouns typically derived from specific words, often nouns associated with individuals' identity. In the case of FAE, it is claimed to originate from an Irish old form of the word *fairy* (Miltersen 2016: 42). Nounself pronouns constitute a distinct class, allowing any noun or word to function as a pronoun based on individual preference. Miltersen (2016: 42) identifies examples like onomatopoeias (*tok, purr*), proper names, and clipped versions of nouns such as *bun/bun/buns/bunself* (from *bunny*) and *bi/bir/birs/birself* (from *bird*). However, it is crucial to note that none of these neopronouns are considered to hold the same status as singular THEY (Hekanaho 2024: 140). Their prominence may result from the rarity of introducing new members to a grammatical paradigm, especially within the context of pronouns being perceived as a closed class resistant to change (Huddleston and Pullum 2002: 425).

Navigating the vast array of neopronouns in use within the nonbinary community poses a considerable challenge, as emphasized by Hakanen (2021: 12), who, while examining XE, ZE, and ZIE in four extensive corpora, retrieved just over one hundred tokens (Hakanen 2021: 14). Given this difficulty, researchers often resort to surveys to elicit pronoun usage (e.g., Hekanaho 2020) or turn to online platforms like forums for data collection (e.g., Zimman 2019). Here, we propose an alternative avenue for exploration: social networks such as *X*, which have proven to be invaluable for investigating authentic language use in the digital sphere (e.g. Tyrkkö *et al.* 2021; Laitinen and Fatemi 2023; Louf *et al.* 2023, to name just a few). Although limited research has delved into NB pronoun usage on *X*, a few related studies have focused on pronoun self-disclosure. Some works reveal disparities and shared patterns among female, male, and nonbinary users (Thelwall *et al.* 2021), while others have gleaned insights into pronoun usage trends (Jiang *et al.* 2022; Tucker and Jones 2023).

These studies yield two primary conclusions: 1) a rising trend in the self-disclosure of gender pronouns on social networks in recent years and 2) the prevalence of SHE as a gender pronoun on *X*, both independently and in combination with others, such as SHE/THEY (Jiang *et al.* 2022; Tucker and Jones 2023). Furthermore, pronoun lists in profiles often intertwine with personal attitudes common among nonbinary *X* users, such as leftist affiliations, the acronym *ACAB* (i.e., *All Cops Are Bastards*), and identifications like *queer*, *trans*, and *pansexual* (Tucker and Jones 2023: 12). Our analysis below will shed more light on these aspects.

3. METHODOLOGY

Established in 2006, the social platform *X*, formerly known as *Twitter*, has evolved into a ubiquitous and influential platform, attracting a diverse user base, including both ordinary individuals and high-profile figures such as celebrities and politicians. With approximately 87 million monthly users in the United States (Semiocast 2023) and a reported usage rate of about 23 percent among U.S. adults (Pew Research 2022a). Thus, *X* has become deeply ingrained in the lives of a significant portion of the population, and this widespread impact positions *X* as a compelling and valuable tool for the scrutiny of human behavior.

X functions as a platform where users can articulate and exchange their ideas, fostering the creation of online conversational threads. Given its nature, *X* provides an ideal environment for investigating the spontaneous production and utilization of language, making it a common choice for linguistic studies (e.g., Zappavigna 2012; Friginal *et al.* 2018; Gonçalves *et al.* 2018; Clarke and Grieve 2019; Grieve *et al.* 2019; Page *et al.* 2022). The platform enables data collection through its Application Programming Interface (API) libraries (Campan *et al.* 2018: 3640). Additionally, analytics platforms like *Followerwonk* (Followerwonk 2022), which offers insights into *X* users, their followers, social authority, and various metrics, facilitate the extraction of valuable information.[7] While *Followerwonk* might not be the most prevalent analytics platform online, scholars have utilized it across different fields of study, ranging from assessing the visibility of financial institutions providing microcredit in Ecuador (Espinoza-Loaiza *et al.* 2017) to exploring pharmaceutical and medical purposes (Styczynski *et al.* 2023). Its versatility in analyzing and extracting meaningful information makes it a valuable tool for collecting data for this study.

The data for this paper was sourced from *X* bios, which are short profiles containing personal information provided by users (this information may include hobbies, place of residence and also icons or emojis). *Followerwonk* was employed for data extraction and the search focused on potential NB pronouns, specifically the nominative forms listed in Table 1, including *they* and others. The search specifically targeted the nominative forms

[7] In 2023, after Elon Musk's acquisition of *Twitter* (and the change of its name to *X*), there have been significant changes in the landscape of *X* analytical platforms, including *Followerwonk*. The platform no longer remains operational with all the functionalities used for this study, as has been acquired by *Fedica* (i.e., https://fedica.com/).

as they represent the unmarked form, which may or may not be accompanied by oblique forms in *X* bios (e.g., *they/them, they/them/their*).

The platform's default presentation of results was organized based on the number of followers for each account. However, the list could be sorted using various metrics, such as the number of tweets, following accounts, account age (measured in days), and social mentions, along with their impact, as illustrated in Figure 1.

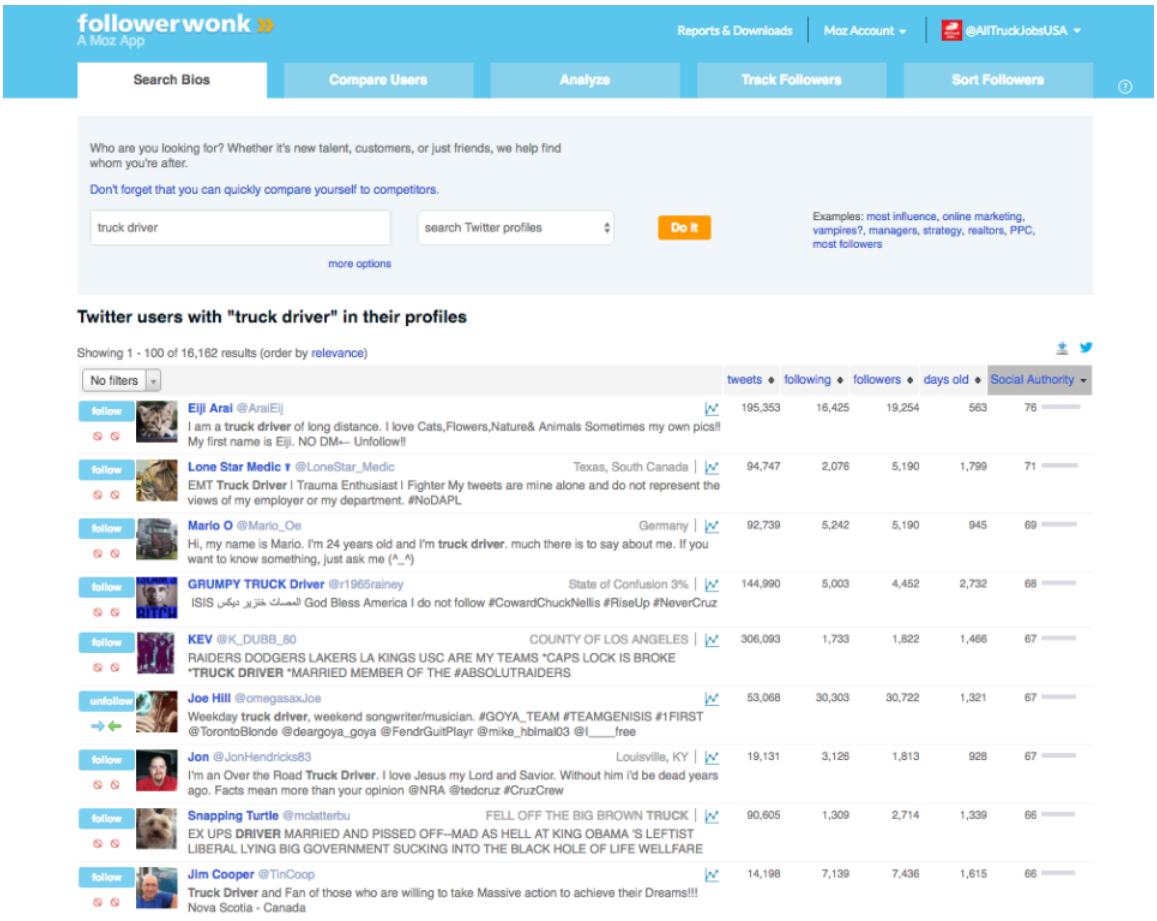

Figure 1: *Followerwonk* results view

The advanced filters provided by *Followerwonk* offer the flexibility to set minimum and/or maximum thresholds for the number of followers, tweets, and following accounts (Figure 2). Notably, the sorting feature by location is a particularly valuable tool. Given that one of the objectives in the study is to examine the potential influence of dominant political views on the choice of NB pronouns in specific regions, we utilized this feature to identify locations with traditions of both Republican and Democrat governments. This information was based on the classification provided by Tausanovitch and Warshaw (2014). The rationale behind the selection of accounts from these particular locations stems from the aforementioned discovery by King and Crowley (2024: 82), who observed

that NBs have played a significant role in shaping online political discussions. They are often perceived as aligning with left-wing ideological positions in the current landscape of US politics and are frequently targeted for ridicule by conservative users of platform *X*.

For the representation of territory with a tendency for liberal governments, New York was chosen as the focal city, because *Followerwonk* allowed us to conduct searches for each of its five boroughs, ensuring an adequate number of tokens for inclusion in our database. On the conservative side, multiple cities were selected. As these cities are not as populous as New York, their results were aggregated to achieve a balanced sample. The chosen cities were specifically identified as standing on the more conservative end of the political spectrum, including Colorado Springs, Fort Worth, Jacksonville, Oklahoma City, Omaha, and, for a larger city example, Miami.

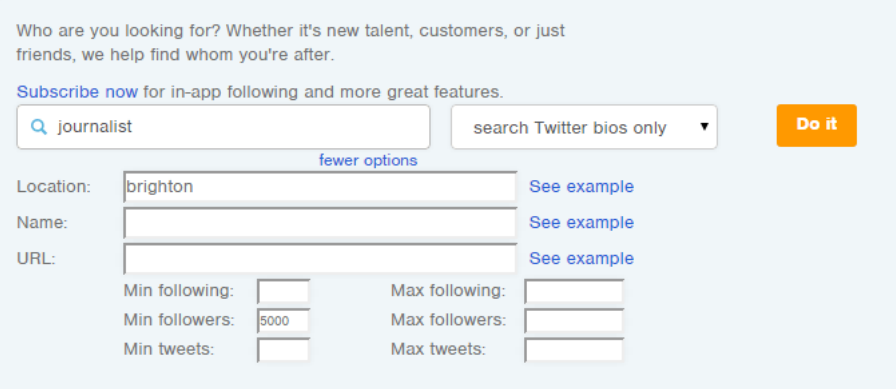

Figure 2: Advanced filters in *Followerwonk*

Thus, each token included in our database was coded for the following extra-linguistic (1–2) and intra-linguistic variables (3–6):

1. **City**: New York, Miami, Colorado Springs, Fort Worth, Jacksonville, Oklahoma City, Omaha. In the case of New York, also **District**: with specification of the five New York districts: Brooklyn, Bronx, Manhattan, Queens, and Staten Island.
2. **Potential activism of the *X* user**: when other ideological keywords (not related to gender) were part of the bio, we noted that in our database (e.g., *climate change*, *Black Lives Matter*, *autism*).
3. **First pronoun mentioned**: In case users resort to rolling pronouns (e.g., *she/they*).

4. **Case** in which the NB pronoun is cited: nominative, accusative and genitive (e.g., *they/them*). Compound forms such as *themselves* are unsuitable candidates to feature in the reduced space allotted to bios in *Twitter*.
5. **Presence of binary pronouns** alongside NB ones: *he* or *she*.
6. **Gender-related keywords** in bios: e.g., *trans, queer, nonbinary, bisexual, cisgender, agender, intersex,* etc.

A comprehensive search using *Followerwonk* identified a total of 12,282 accounts featuring NB pronouns within the explored territories. Specifically, there were 6,432 accounts in New York and 5,850 in the other cities (with a tradition of Republican governments). From each group, a sample of approximately 1,000 accounts was systematically chosen by adjusting the sorting options of the analytics platform. That is, since the 12,282 accounts could not be downloaded from *Followerwonk* for randomization, the only feasible approach to selecting a somewhat random sample was to sort the accounts based on factors such as account age and social authority. These factors were deemed to have a negligible impact on the use of NB pronouns and were thus not expected to introduce bias into the results. As the summarized results presented in Tables 2 and 3 show, a total of 1,980 accounts were analyzed.

	Total number of accounts with NB pronouns	**Accounts selected**
Bronx	776	151
Brooklyn	3,502	418
Manhattan	955	176
Queens	1,038	230
Staten Island	161	37
Total	**6,432**	**1,012**

Table 2: Number of accounts collected from New York boroughs and the total number of results in *Followerwonk*

	Total number of accounts with NB pronouns	**Accounts selected**
Colorado Springs	446	122
Fort Worth	717	131
Jacksonville	811	122
Miami	2,652	296
Oklahoma City	606	131
Omaha	618	166
Total	**5,850**	**968**

Table 3: Number of accounts collected from US cities with a tradition of Republican governments and the total number of results in *Followerwonk*

4. RESULTS

4.1. Monopronouns and rolling pronouns

X users have the option of self-definition through a single pronoun (e.g., *they*), termed 'monopronoun' use, or a combination of pronouns (e.g., *they/he, she/they/xe*), known as 'rolling pronouns' —defined as "the use of multiple pronouns that can be used alternately or shift over time" (LGBTQ Nation 2022). Interestingly, rolling pronouns emerge as the prevailing trend in our dataset: as illustrated in Table 4, 65 percent of the scrutinized accounts in New York opt for multiple pronouns to articulate their gender identity, while 35 percent identify as monopronoun users. This statistically significant difference[8] also holds for the other cities (65.6% of accounts exhibiting rolling pronouns *vs.* 34.4% of accounts showing monopronouns), suggesting a consistent pattern of pronoun usage.

	New York accounts	**Other cities accounts**	**Total**
Monopronouns users	354 (35%)	333 (34.4%)	687 (34.7%)
Rolling pronouns users	658 (65%)	635 (65.6%)	1,293 (65.3%)
Total	**1,012**	**968**	**1,980**

Table 4: Monopronoun and rolling pronoun users by community

The choice between a monopronoun and rolling pronouns significantly impacts the prevalence of inflectional forms other than the nominative in our dataset. Notably, monopronouns are frequently accompanied by non-nominative forms (98%),[9] while rolling pronouns exhibit a lower proportion in this regard (90%), as illustrated in Table 5. This significant[10] contrast between monopronouns and rolling pronouns can be attributed, in part, to the character limit (160) imposed on bios in *X*. Users employing rolling pronouns often prioritize conciseness due to character constraints, limiting the inclusion of additional inflectional forms in favor of other aspects of their personal profile. Nevertheless, ten percent of rolling pronoun users do include additional forms, as exemplified by constructions such as (i) *he/him they/them she/her* or (ii) *they/them xe/xem*. In contrast, monopronoun users predominantly opt for the nominative/accusative form, potentially reflecting a formulaic expression signaling the use of pronouns for

[8] The test applied to these data is the Z score test, which calculates the value of *z* (and associated *p* value) for two population proportions. This test compares the observed frequency with the expected frequency; the *z* score is the number of standard deviations from the mean frequency, in such a way that the higher the *z* score, the lower the likelihood that only chance is affecting the distribution (McEnery et al. 2006: 57). In this case, the value of *z* is 19.2599. The value of *p* is < .00001. The result is significant at p < .05 (https://www.socscistatistics.com/tests/ztest/default.aspx)

[9] On most cases, the non-nominative form is in the accusative, because the genitive form has shown to be anecdotal with only 39 cases from almost 2,000 tokens from the dataset.

[10] The value of *z* is 35.029. The value of *p* is < .00001. The result is significant at p < .05.

gender identity purposes rather than as components of a broader linguistic structure. In essence, constructions like *they/them, ze/zim/zer* have become conventional ways of conveying one's preferred pronouns (pronouns.org 2023).

	Only nominative form	**Other inflectional forms**	**Total**
Monopronouns users	14 (2%)	673 (98%)	687
Rolling pronouns users	1,164 (90%)	129 (10%)	1,293
Total	**1,178 (59.5%)**	**802 (40.5%)**	**1,980**

Table 5: Presence of inflectional forms other than the nominative with monopronouns and rolling pronouns

An important finding in our analysis is that monopronoun users overwhelmingly favor singular *they* (Table 6). Specifically, only 21 accounts opt for a single neopronoun, with nine choosing *ze*, nine selecting *xe*, and three opting for *ey* —each accompanied by distinct non-nominative forms. All other neopronouns examined in this study are found within rolling pronouns, predominantly led by *they* (49%). Following this are *she* (29.5%), *he* (20.4%), and neopronouns collectively, constituting a mere 1.1% of all accounts with rolling pronouns. Notably, there are minimal discrepancies between territories, with *they* being more frequent in New York than in the other cities analyzed (51.7% vs. 46.3%). Conversely, *she* exhibits a higher frequency in other cities (32.1% vs. 27%), as shown in Table 6:

First chosen pronoun	**New York**	**Other cities**	**Total**
THEY	523 (51.7%)	448 (46.3%)	971 (49%)
HE	205 (20.3%)	199 (20.6%)	404 (20.4%)
SHE	273 (27%)	311 (32.1%)	584 (29.5%)
Neopronouns	11 (1%)	10 (1%)	21 (1.1%)
Total	**1,012**	**968**	**1,980**

Table 6: First pronoun chosen by *X* users in rolling pronouns: New York vs. other cities

The incorporation of gendered pronouns alongside NB pronouns is a prevalent phenomenon in our dataset, since a total of 1,254 accounts feature either *he*, *she*, or a combination of both, as illustrated in Table 7:[11]

	Raw Frequency	**Percentage**
HE	482	38.4%
SHE	712	56.8%
HE AND SHE	60	4.8%
Total	**1,254**	**100%**

Table 7: Nonbinary users in our dataset with at least one gendered pronoun

[11] Out of these 1,254 accounts that list gendered pronouns alongside NB ones, 28 also list neopronouns, while 1,226 only list THEY and SHE, HE or HE and SHE.

Furthermore, our analysis of *X* accounts reveals that when users opt for gendered pronouns alongside NB pronouns, they predominantly choose *he* or *she* as their first pronoun before specifying their NB pronoun. Table 8 illustrates this trend, indicating that 74.3 percent of users prefer HE (e.g., *he/they*), mirroring the 75.5 percent of users who opt for SHE (e.g., *she/they*).

First chosen pronoun	Bios with HE	Bios with SHE	TOTAL
HE	403 (74.3%)	27 (3.5%)	430
SHE	14 (2.6%)	583 (75.5%)	597
THEY	123 (22.7%)	161 (20.9%)	284
Neopronouns	2 (0.4%)	1 (0.1%)	3
Total	**542**	**772**	**1,314**

Table 8: First pronoun in set of rolling pronouns that include (binary) gendered HE and SHE

Table 8 also highlights the infrequent occurrence of neopronouns within rolling pronouns that include gendered *he* or *she*. However, a comprehensive examination of the entire set of rolling pronoun options in the analyzed accounts reveals that neopronouns are not uncommonly selected as second or subsequent options by *X* users: For instance, examples such as 1) *they (ey/em/eir)*, 2) *she/he/they/xe/xim*, and the most elaborate instance in our dataset, 3) *he/ him /his /she /her /sher /hershey's /zhe/zher /zir/xyr/they/them/thems/they're/their/there/thon/fae/I/me/you/your/you're/us/y'all/we/ wumbo/it/that/this/thit/pronoun*. The specific frequency of neopronouns in comparison to *they* is detailed in Section 4.2 below.

4.2. Type of NB pronoun: THEY and neopronouns

Table 9 shows the frequency of all NB pronouns found. The data clearly shows the prevalence of *they* (95% of all cases). As mentioned, these pronouns could appear in any position in the users' bios, since neopronouns hardly ever appear as monopronouns and, for that reason, the total number of tokens surpasses the number of accounts analyzed.

	New York	Other cities	Total
They	982	952	1,934 (95.%)
Nounself pronoun	18	5	23 (1.1%)
Xe	10	10	20 (1%)
It	9	8	17 (0.9%)
Ze	11	2	13 (0.6%)
Foreign pronouns	8	4	12 (0.6%)
Fae	6	1	7 (0.3%)
Any (pronoun)	2	3	5 (0.3%)
Ey	2	2	4 (0.2%)
Total	**1,048**	**987**	**2,035**

Table 9: Distribution of NB pronouns in the dataset

In addition to reinforcing the nonbinary status of *they*, Table 9 also arranges the neopronouns from Table 1 as follows: nounself pronouns exhibit the highest prevalence (23 tokens in total, e.g., *pup* or *neigh*), followed by *xe* (20 tokens), *ze* (13), *fae* (7), and *ey* (4). Furthermore, Table 1 includes other pronominal forms discovered incidentally (as a second or later option in rolling pronouns). These include the pronoun *it* (17), foreign pronouns such as *elle* (from Spanish)[12] or *sie* (from German) (12), as well as *any*, a concise form standing for *any pronoun* (5), suggesting a clear flexibility in the users' choice of pronouns. The presence of nounself pronouns is noteworthy, considering their diverse nature, with almost none repeated (e.g., *thude*, *neon*, or *bruh*; exemplified in *he/they/neigh/bruh/skull/neon*), except for *fae*, which occurs several times and could be included in this category. The NB pronoun *it* also appears with relative frequency, despite assertions that it may be dehumanizing and perilous (Norris and Welch 2020: 9). Some users express comfort with being referred to with this pronoun alongside other NB pronouns (e.g., *they/it*; *they/it/ze*; or *xe/they they/jze/it*). Additionally, *X* users have incorporated NB pronouns from other languages, such as *elle*, proposed in Spanish, and *sie*, representing the third person singular feminine and also the plural in German (e.g., *he/they El/Elle; they/sie/them*), serving as anecdotal evidence of the multilingual nature of the social network, despite its overwhelming English-speaking majority (Grandjean 2016: 6). Regarding differences between political territories, due to the overall small number of neopronouns, no statistical test can be applied, and the distinctions between territories traditionally ruled by Democrats and Republicans do not seem to be relevant.

4.3. Keywords in bios

The final result concerning the variables in our dataset (outlined in Section 3 above) pertains to the presence of lexical keywords in *X* bios related to gender identity, sexuality (e.g., *queer*, *trans*, *bisexual*), or various forms of activism related to different causes (e.g., *climate change*, *Black Lives Matter*, *autism*). After the manual examination of the 1,980 accounts analyzed, the findings indicate that 26.7 percent of *X* accounts (n= 529) incorporate keywords reflecting their sexual or gender identity (as depicted in Figure 3),

[12] The pronoun *elle* is often listed as a NB in Spanish (e.g. López 2019), which has led us to consider this a NB in this context (instead of the homograph French feminine pronoun).

whereas the inclusion of personal and political keywords is slightly lower, accounting for 13.6 percent (n= 270, as illustrated in Figure 4).

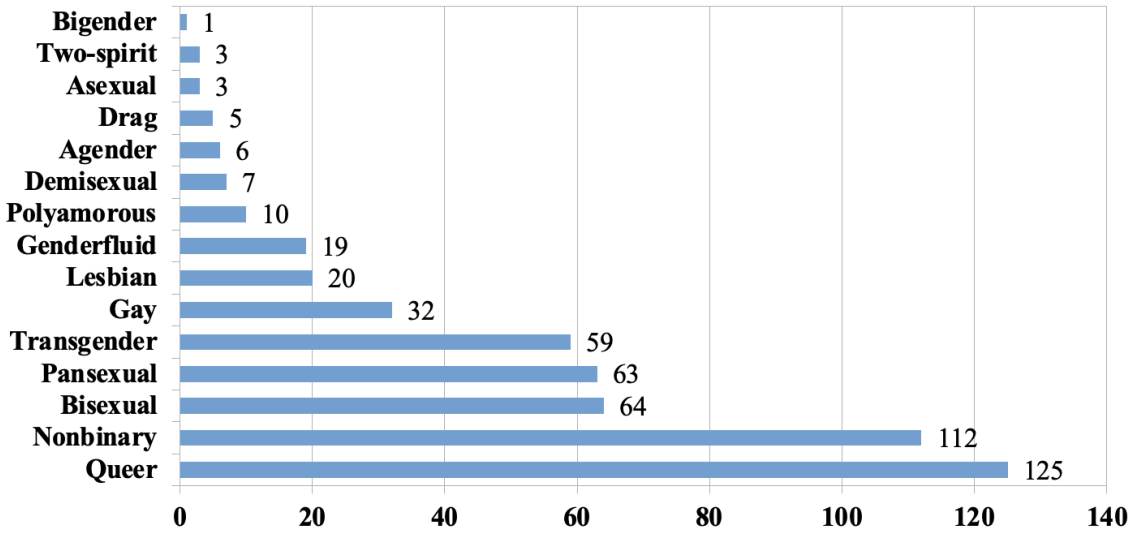

Figure 3: Keywords related to gender identity or sexuality that accompany NB pronouns in our dataset

Figure 3 provides an overview of the frequency of keywords associated with gender identity and sexuality that co-occur with NB pronouns in our dataset. The observed keywords can be categorized into four primary blocks. Firstly, *queer* emerges as the most prevalent term in *X* profiles, appearing 125 times, closely followed by *nonbinary* with 112 instances. In the second block, the triad of *bisexual* (64), *pansexual* (63), and *transgender* (59) takes precedence. The third block comprises terms such as *gay* (32), *lesbian* (20), and *genderfluid* (19). Lastly, we encounter less frequent terms like *polyamorous* (10), *demisexual* (7), *agender* (6), *drag* (5), *asexual* (3), *two-spirit* (3), a characteristic term within the Native American community and *bigender* (1).

Figure 4 highlights the prevalence of additional keywords in our dataset that offer insights into users' profiles. At the forefront is the acronym *BLM*, which stands for *Black Lives Matter*, appearing in 120 accounts. Following closely is another acronym, *NSFW* (*Not Suitable/Safe for Work*), present in 90 accounts, often associated with explicit or inappropriate material rather than specific political affiliations. In the third and fourth positions, we encounter terms that bring visibility to minority groups: 9 instances of *neurodivergent* and 13 of *disabled*. The list continues with politically charged labels, including *ACAB* (*All Cops Are Bastards*) in six accounts, *Free Palestine* in four, and three instances each of *Pan-Africanism* and *Feminist*, and two of *Abolitionist*. The significance of these figures lies more in their qualitative implications than their quantitative

representation. As demonstrated in prior studies (e.g., Tucker and Jones 2021), data hint at a connection between actively articulating one's nonbinary identity and expressing overt support for specific social causes.

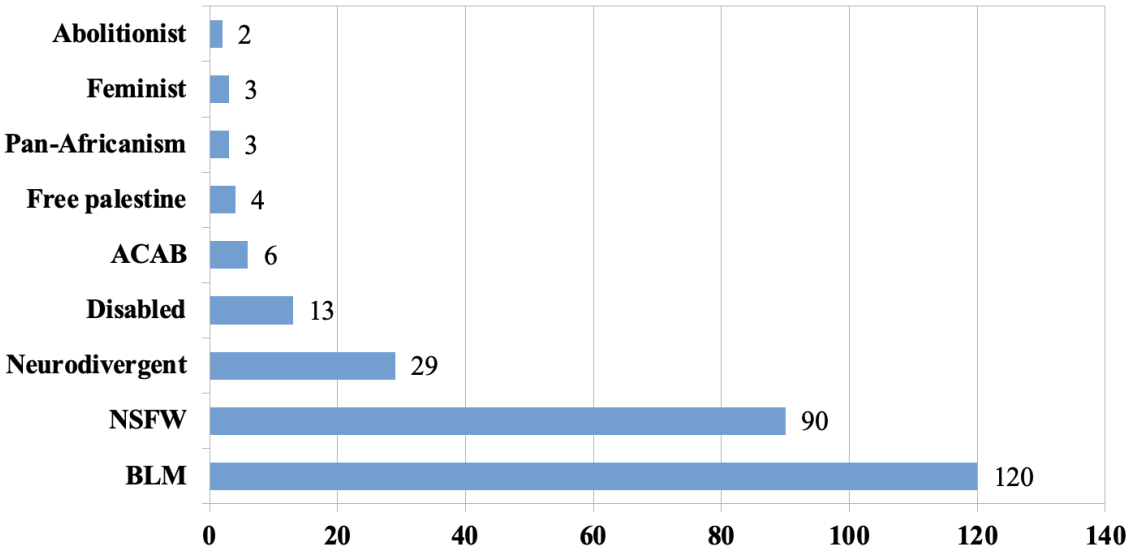

Figure 4: Keywords related to some kind of activism on *X* that accompany NB pronouns in our dataset

5. Discussion

This paper has studied the presence of NB pronouns in *X* profiles, with the aim of determining the factors that might condition the variation among the myriad of NB pronouns available as of 2023 (see Table 1). One such factor was considered to be the place of residence of *X* users, and for that reason data were collected (using the extinct *X* analytics platform *Followerwonk*) based on geographical or political factors. Two samples were taken from a city traditionally ruled by Democrats, namely New York, and several cities traditionally ruled by Republicans. The results do not conclusively establish a correlation between political affiliations of a territory and pronoun choices by the citizens. Thus, our results show no significant differences between users in both kinds of territory regarding aspects such as the frequency of monopronouns and rolling pronouns (Table 4), the pronoun that occupies first position in rolling pronouns (Table 6), or the particular frequency of THEY and the neopronouns (Table 9). This can very well be interpreted as the result of the global character of online communities, which tend to behave alike regardless of their particular geographical location, as has been previously found for K-pop communities (Malik and Haidar 2020: 11). Thus, although notable differences have been found in previous literature between the use of *X* by Republicans

(17%) and Democrats (32%) (Pew Research 2022b), one cannot conclude either that i) all *X* users living in a city ruled by one party follow their political views, or that ii) the main political view of a geographical territory is the only influence on netizens in an increasingly globalized word. Therefore, it looks as if the once claimed true democratic nature of social networks (e.g., Orr *et al.* 2009), where everyone had a voice and social differences were erased is still at work among NB individuals on *X*.

The unequivocal dominance of the pronoun THEY emerges as a defining characteristic within the dataset. This overwhelming usage (1,953 out of 1,980 accounts) supports the argument that THEY is the most widely accepted NB pronoun, overshadowing neopronouns in popularity (also noted by Hekanaho 2020: 222). The closed nature of the pronoun system, where new forms like neopronouns struggle for acceptance, contrasts with the smoother transition provided by THEY, which despite having been proscribed in usage guides for over two centuries has found its way into standard varieties of English very much thanks to the non-sexist language reform initiated by second-wave feminism in the 1960s (Paterson 2020: 261–264). Thus, in the battle for non-sexist language feminists defended the use of singular THEY or combined HE OR SHE and both were consistently neglected by the gate-keepers of the language, on the basis that the former violates number agreement with its antecedent and the latter leads to a cumbersome style. In the twenty-first century, however, and among nonbinary individuals, the otherwise proscribed THEY is considered as "more reasonable" than the neopronouns (Hekanaho 2020: 222), as it is seen as more familiar and easier to educate family and friends on the reference towards nonbinary individuals (McGlashan and Fitzpatrick 2018: 12; Cordoba 2020: 58). Among neopronouns, according to our results, nounself pronouns head the list, on most occasions with nonce forms such as THUDE, NEON or BRUH, and they are followed by XE, IT, ZE, foreign pronouns, FAE, ANY and, finally EY (see Table 9). The multiplicity of options available reveals i) that linguistic creativity has no boundaries, ii) that gender identity is very complex and multifaceted and individuals enjoy the possibility of choosing how they want to be referred to, and iii) that we may be in the midst of a case of language variation that will end up in the survival of one or several pronominal forms if such forms manage to seamlessly integrate into the linguistic paradigm. The higher their integration, the higher their accessibility for individuals outside the LGBTQI+ community, and among these THEY is said to be clear winner (Hekanaho 2020: 222), as our results support.

Despite this preference for THEY, we have also seen that a vast majority of *X* users define their identity by rolling pronouns, highlighting a preference for multiple pronouns over a single one. This term encompasses individuals who may alter their pronouns based on context or employ them regularly, indicating the fluidity of gender identity expression. The prevalence of rolling pronouns users may be attributed to factors like gender fluidity or the comfort nonbinary individuals feel using multiple pronouns during transitional phases (McGlashan and Fitzpatrick 2018: 9; Jiang *et al.* 2022). This is in fact supported by the fact that gendered pronouns exhibit a much higher frequency than expected (1,254 tokens in our dataset include either HE OR SHE in the list of pronouns of choice alongside other NB pronouns). However, we acknowledge that more qualitative investigation will be necessary to understand specific preferences in different contexts.

The analysis of rolling pronouns in *X* bios also revealed that inflectional forms other than the nominative tend to be absent (90% of the times, as seen in Table 5), while it is overwhelmingly present in monopronouns (98%). A potential explanation for the absence of oblique forms is the 160-character limit in *X* bios, but that does not explain its practically total presence in the case of monopronouns. In that case, we believe that the near-formulaic nature of the combination of nominate and accusative or genitive forms (e.g., *they/them* or *they/them/their*) constitutes a well-established linguistic chunk associated with the communication of gender identity.

As expected, the use of NB pronouns correlates largely with the presence of lexical terms related to gender and sexuality (Figure 3). Likewise, political ideologies and personal beliefs find expression on *X*, with left-wing ideologies prominently represented through keywords like *BLM* and *ACAB* (as already mentioned by Tucker and Jones 2023: 11). Our results list these and other politically oriented key terms (Figure 4) and also highlights the inclusion of *NSFW* as a prevalent keyword, which suggests a shift in online discourse, reflecting a growing inclusion of explicit content. Additionally, the emergence of keywords related to neurodivergence, such as *autistic*, aligns with the notion that certain nonbinary individuals may have a higher likelihood of being neurodivergent (McClurg 2023). This intersectionality hints at the complex interplay between gender identity and neurodiversity, urging further exploration within this intersection.

6. CONCLUSIONS

Transforming English into a more inclusive language is a challenging task and nonbinary individuals find on social networks, such as *X*, a way of expressing their identity freely. A key strategy for claiming identity involves the selection of personal pronouns. This paper has contributed to the ongoing discourse on NB pronouns by scrutinizing the pronouns chosen by users in 1,980 *X* accounts. The analysis aimed to uncover sociolinguistic patterns among the myriad of NB pronouns available, considering both extra-linguistic and intra-linguistic variables. In the examination of extra-linguistic variables, we have scrutinized the role played by municipal political government, reflecting the overall Democrat or Republican majority in various cities. Additionally, we assessed the potential activism of users by considering the presence of lexical keywords related to specific political issues. Within intra-linguistic variables, we examined firstly the order of pronouns, particularly in cases where more than one pronoun was chosen — a prevalent occurrence in 65.3 percent of all cases, exemplified by rolling pronouns like *they/xe*. Secondly, we investigated the presence of inflectional forms beyond the nominative, such as *they/them*. Finally, the analysis also encompassed the presence of binary gendered pronouns, *he* and/or *she*, and the selection of lexical gender-related vocabulary within the *X* bio.

Through a meticulous examination of 1,980 *X* accounts, a distinct pattern emerged, overwhelmingly favoring the use of *they* among nonbinary users, evident in 1,953 instances (RQ1). This prevalence constitutes a case of (quasi-)standardization, challenging traditional proscriptions that survived until the twenty-first century (as an example, Batko's 2004 usage guide still considers singular THEY a mistake when used with singular antecedents such as *everyone*). Beyond *they*, the dataset reveals the presence of other NB pronouns, frequently embedded in rolling pronouns. Nounself pronouns (Miltersen 2016), including THUDE, NEON, and BRUH, take the lead, followed closely by XE, IT, ZE, foreign pronouns, FAE, ANY, and, ultimately, EY. Despite this diversity, all neopronouns collectively constitute only five percent of the entire set of NB pronouns in our dataset (Table 9). This observation suggests that the path paved by feminists in the non-sexist language reform has predominantly favored the acceptance of singular THEY, a usage that has persisted since medieval times.

The prevalence of THEY, however, coexists with the utilization of (binary) gendered pronouns (HE and/or SHE), collectively appearing on 1,254 occasions (RQ2) within the

context of rolling pronouns. This co-occurrence suggests a transitional phase for some individuals, as tentatively interpreted in line with McGlashan and Fitzpatrick (2018: 9) and Jiang *et al.* (2022).

The distinction between rolling pronouns and monopronouns significantly influences the presence of inflectional forms beyond the nominative (RQ3). While rolling pronouns predominantly manifest in the nominative form in 90 percent of instances, monopronouns exhibit an oblique form 98 percent of the times. This discrepancy is interpreted as a consequence of the formulaic nature of the nominative/oblique form of the pronoun, showcasing a conventionalized way of expressing one's identity.

The political traditions of the cities where the *X* users reside (RQ4) has proven to be a non-significant factor in explaining the variation among NB pronouns. This uniform behavior exhibited by *X* users, irrespective of territorial factors, is attributed to the difference-erasing role of social networks. Profiles tend to conform more with the globalized nature of the internet than with specific geographical neighbors.

Addressing RQ5, our work reveals a remarkable correlation between the presence of NB pronouns and lexical keywords related to gender and sexuality on one hand, and political activism on the other. This correlation suggests that individuals on the social network utilize NB pronouns as part of a broader strategy for activist purposes, aligning with a trend to increase visibility and assert their rights as citizens.

In conclusion, the comprehensive analysis of NB pronoun usage on *X* offers valuable insights into the intricate connections between language, identity, and online dynamics. The dominance of THEY, the emergence of rolling pronouns users, and the challenges faced by neopronouns underscore the nuanced nature of gender identity expression in digital spaces. Our study is subject to certain limitations, including the restricted sample size of *X* accounts examined, the potential bias introduced by *Followerwonk*, and the focus solely on US-based accounts. Consequently, it is important to refrain from interpreting our findings as indicative of the global English-speaking community's perspectives on *X*. Instead, they should be regarded as a gateway to further exploration of online spaces. Thus, other avenues should be explored, like the intersectionality of gender identity, political expressions, and linguistic choices, providing a rich foundation for future research within the LGBTQI+ community.

REFERENCES

Balhorn, Mark. 2009. The epicene pronoun in contemporary newspaper prose. *American Speech* 84/4: 391–413.

Baron, Dennis. 2010. *The Gender-Neutral Pronoun: 150 Years Later, still an Epic Fail. OUPblog.* https://blog.oup.com/2010/08/gender-neutral-pronoun/ (25 June, 2023.)

Baron, Dennis. 2018. A Brief History of Singular 'They'. *The Web of Language. Dennis Baron's Go-to Site for Language and Technology in the News.* https://www.oed.com/discover/a-brief-history-of-singular-they/?tl=true (27 April, 2024.)

Baron, Dennis. 2020. *What's Your Pronoun? Beyond He and She*. London: Liveright.

Batko, Ann. 2004. *When Bad Grammar Happens to Good People. How to Avoid Common Errors in English*. Franklin Lakes: Career Press.

Bodine, Ann. 1975. Androcentrism in prescriptive grammar: Singular 'they', sex-indefinite 'he', and 'he or she'. *Language in Society* 4: 129–146.

Bradley, Evan. 2020. The influence of linguistic and social attitudes on grammaticality judgments of singular 'they'. *Language Sciences* 78: 1–11.

Bradley, Evan, Julia Salkind, Ally Moore and Sofi Teitsort. 2019. Singular 'they' and novel pronouns: Gender-neutral, nonbinary, or both? *Proceedings of the Linguistic Society of America* 4/36: 1–7.

Bucholtz, Mary and Kira Hall. 2010. Locating identity in language. In Carmen Llamas and Dominic James Landon Watt eds. *Language and Identities*. Edinburgh: Edinburgh University Press, 18–28.

Campan, Alina, Tobel Atnafu, Traian Marius Truta and Joseph Nolan. 2018. Is data collection through Twitter reaming API useful for academic research? *IEEE International Conference on Big Data*: 3638–3643. https://doi.org/10.1109/BigData.2018.8621898.

Clarke, Isobelle and Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE* 14/9: e0222062. https://doi.org/10.1371/journal.pone.0222062.

Conrod, Kirby. 2019. *Pronouns Raising and Emerging*. Washington: University of Washington doctoral dissertation.

Cordoba, Sebastian. 2020. *Exploring Non-Binary Genders: Language and Identity*. Leicester: De Montfort University dissertation.

Curzan, Anne. 2003. *Gender Shifts in the History of English*. Cambridge: Cambridge University Press.

Dembroff, Robin and Daniel Wodak. 2018. He/she/they/ze. *Ergo: An Open Access Journal of Philosophy* 5/14: 371–406.

Espinoza-Loaiza, Viviana, Rosario Puertas, Valentin Martínez, Aurora Samaniego-Namicela and Eulalia-Elizabeth Salas-Tenesaca. 2017. Visibility and impact of the microcredit and the digital social media: A case study of financial institutions in Ecuador. In Francisco Campos ed. *Media and Metamedia Management*. Switzerland: Springer, 413–418.

Followerwonk. 2022. *Tools for Twitter Analytics, Bio Search and More*. https://followerwonk.com/search-bio.html (30 December, 2022.)

Friginal, Eric, Oksana Waugh and Ashley Titak. 2018. Linguistic variation in Facebook and Twitter posts. In Eric Friginal ed. *Studies in Corpus-Based Sociolinguistics*. New York: Routledge, 342–363.

Gender Census. 2022. *Gender Census 2022: Worldwide Report.* https://www.gendercensus.com/results/2022-worldwide/#pronouns (25 June, 2023.)

Gonçalves, Bruno, Lucía Loureiro-Porto, José J. Ramasco and David Sánchez. 2018. Mapping the americanization of English in space and time. *PLoS ONE* 13/5: e0197741. https://doi.org/10.1371/journal.pone.0197741

Grandjean, Martin. 2016. A social network analysis of Twitter: Mapping the digital humanities community. *Cogent: Arts & Humanities* 3. https://doi.org/10.1080/23311983.2016.1171458.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami and Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2/11. https://doi.org/10.3389/frai.2019.00011.

Hakanen, Lotta. 2021. "*Let xir do Whatever xe wants": A Corpus Study on Neopronouns ze, xie and zie*. Tampere: Tampere University Bachelor's Thesis.

Harmon, Amy. 2019. 'They' is the Word of the Year, Merriam-Webster Says, Noting its Singular Rise. *New York Times*. https://www.nytimes.com/2019/12/10/us/merriam-webster-they-word-year.html#:~:text=Merriam%2DWebster%20announced%20the%20pronoun,whose%20gender%20identity%20is%20nonbinary (1 July, 2023.)

Hegarty, Peter, Y. Gavriel Ansara and Meg-John Barker. 2018. Nonbinary gender identities. In Nancy K. Dess, Jeanne Marecek and Leslie C. Bell eds. *Gender, Sex, and Sexualities: Psychological Perspectives*. New York: Oxford University Press: 53–76.

Hekanaho, Laura. 2020. *Generic and Nonbinary Pronouns: Usage, Acceptability and Attitudes*. Helsinki: University of Helsinki dissertation.

Hekanaho, Laura. 2024. The communicative functions of 3rd person singular pronouns: Cisgender and transgender perspectives. In Minna Nevala and Minna Palander-Collin eds. *Self- and Other-Reference in Social Contexts. From Global to Local Discourses*. Amsterdam: John Benjamins, 138–165.

Herdt, Gilbert. 1996. *Third Sex, Third Gender: Beyond Sexual Dimorphism in Culture and History*. Brooklyn: Zone Books.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ingram, David. 2023. *Elon Musk's New Twitter Pronoun Rule Invites Bullying, LGBTQ Groups Say*. *NBC News*. 2 June 2023. https://www.nbcnews.com/tech/tech-news/elon-musks-new-twitter-pronoun-rule-invites-bullying-lgbtq-groups-say-rcna87336 (30 September, 2023.)

Jiang, Julie, Emily I. Chen, Luca Luceri, Goran Muric, Francesco Pierri, Ho-Chun Herbert Chang, and Emilio Ferrara. 2022. What are your pronouns? Examining gender pronoun usage on Twitter. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2207.10894.

King, Brian W. and Archie Crowley. 2024. The future of pronouns in the online/offline nexus. In Laura Louise Paterson ed., 74–86.

Konnelly, Lex and Elizabeth Cowper. 2020. Gender diversity and morphosyntax: An account of singular *they*. *Glossa: A Journal of General Linguistics* 5/1: 40. https://doi.org/10.5334/gjgl.1000.

Konnelly, Lex, Kirby Conrod and Evan D. Bradley. 2024. Non-binary singular *they*. In Laura Louise Paterson ed., 450–464.

Laitinen, Mikko. 2024. A history of personal pronouns in Standard English. In Laura Louise Paterson ed., 29–43.

Laitinen, Mikko and Masoud Fatemi. 2023. Data-intensive sociolinguistics using social media. *Annales Academiae Scientiarum Fennicae* 2023/2: 38–61.
LaScotte, Darren K. 2016. Singular *they*: An empirical study of generic pronoun use. *American Speech* 91/1: 62–80.
LGBTQ Nation. 2022. *Why some People Use She/They & He/They Pronouns. LGBTQ Nation.* https://www.lgbtqnation.com/2022/05/people-use-pronouns/ (25 June, 2023.)
Lindqvist, Anna, Emma Aurora Renström and Marie Gustafsson Sendén. 2019. Reducing a male bias in language? Establishing the efficiency of three gender-fair language strategies. *Sex Roles* 81: 109–117.
López, Ártemis. 2019. Tú, yo, elle y el lenguaje no binario. *La Linterna del Traductor* 19: 142–150.
Louf, Thomas, Bruno Gonçalves, José J. Ramasco, David Sánchez and Jack Grieve. 2023. American cultural regions mapped through the lexical analysis of social media. *Humanities and Social Sciences Communications* 10. http://dx.doi.org/10.1057/s41599-023-01611-3
Loureiro-Porto, Lucía. 2020. (Un)democratic epicene pronouns in Asian Englishes: A register approach. *Journal of English Linguistics* 48/3: 282–313.
Lund Eide, Mari. 2018. *Shaping the Discourse of Gender-Neutral Pronouns in English: A Study of Attitudes and Use in Australia*. Bergen: University of Bergen Master's Thesis.
Malik, Zunera and Sham Haidar. 2020. Online community development through social interaction – K-pop stan Twitter as a community of practice. *Interactive Learning Environments* 31/2: 1–19.
Matsuno, Emmie and Stephanie L. Budge. 2017. Non-binary/genderqueer identities: A critical review of the literature. *Current Sexual Health Reports* 9: 116–120.
McClurg, Lesley. 2023. *Transgender and Nonbinary People Are up to Six Times More Likely to Have Autism*. *NPR Hour Program Stream*. https://www.npr.org/2023/01/15/1149318664/transgender-and-non-binary-people-are-up-to-six-times-more-likely-to-have-autism (26 June, 2023.)
McEnery, Tony, Richard Xiao and Yukio Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.
McGlashan, Hayley and Katie Fitzpatrick. 2018. I use any pronouns, and I'm questioning everything else: Transgender youth and the issue of gender pronouns. *Sex Education* 18: 239–252.
McLemore, Kevin. 2015. Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity* 14/1: 51–74..
Merriam-Webster. 2019. *A Note on the Nonbinary 'They'*. *Merriam-Webster*. https://www.merriam-webster.com/words-at-play/nonbinary-they-is-in-the-dictionary (20 November, 2023.)
Miltersen, Ehm Hjorth. 2016. Nounself pronouns: 3rd person personal pronouns as identity expression. *Journal of Language Works* 1/1: 37–62.
Norris, Marcos and Andrew Welch. 2020. Gender pronoun use in the university classroom: A post-humanist perspective. *Transformation in Higher Education* 5/0: a79. https://doi.org/10.4102/the.v5i0.79.
Orr, Emily S., Mia Sisic, Craig Ross, Mary G. Simmering, Jaime M. Arseneault and R. Robert Orr. 2009. The influence of shyness on the use of Facebook in an undergraduate sample. *CyberPsychology & Behavior* 12/3: 337–340.
Page, Ruth, David Barton, Carmen Lee, Johann Wolfgang Unger and Michele Zappavigna. 2022, *Researching Language and Social Media*. London: Routledge.

Parker, Linden. 2017. *An Exploration of Use of and Attitudes towards Gender-Neutral Pronouns among the Non-Binary, Transgender and LGBT+ Communities in the United Kingdom*. Colchester: University of Essex Master's Thesis.

Paterson, Laura Louise. 2011. Epicene pronouns in UK national newspapers: A diachronic study. *ICAME Journal* 35: 171–184.

Paterson, Laura Louise. 2014. *British Pronoun Use, Prescription, and Processing. Linguistic and Social Influences Affecting "they" and "he."* Basingstoke: Palgrave.

Paterson, Laura Louise. 2020. Non-sexist language policy and the rise (and fall?) of combined pronouns in British and American written English. *Journal of English Linguistics* 48/3: 258–281.

Paterson, Laura Louise ed. 2024. *The Routledge Handbook of Pronouns*. New York: Routledge.

Pew Research. 2022a. *Social Media and News Fact Sheet. Pew Research Center.* https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/ (21 June, 2023.)

Pew Research. 2022b. *10 Facts about Americans and Twitter. Pew Research Center.* https://www.pewresearch.org/short-reads/2022/05/05/10-facts-about-americans-and-twitter/ (7 June, 2023.)

Pronouns.org. 2023. *Resources on Personal Pronouns*. https://pronouns.org/ (20 November, 2023.)

Pronouns.page. 2024. *List or Popular Pronouns*. https://en.pronouns.page/pronouns/#google_vignette (27 April, 2024.)

Pronouny. 2016–2020. *Public Pronoun List*. https://pronouny.xyz/pronouns/list/public (27 April, 2024.)

RAE (Real Academia de la Lengua Española). 2023. *Los Sinónimos y Antónimos se Incorporan al «Diccionario de la Lengua Española» en su Actualización 23.7.* https://www.rae.es/noticia/los-sinonimos-y-antonimos-se-incorporan-al-diccionario-de-la-lengua-espanola-en-su (28 November, 2023.)

Scelfo, Julie. 2015. *A University Recognizes a Third Gender: Neutral*. *New York Times*. http://www.nytimes.com/2015/02/08/education/edlife/a-university-recognizes-a-third-gender-neutral.html (8 June, 2023.)

Semiocast. 2023. *Number of Twitter Users by Country. Semiocast.* https://semiocast.com/number-of-twitter-users-by-country/ (21 June, 2023.)

Simpson, Lauren and Jean-Marc Dewaele. 2019. Self-misgendering among multilingual transgender speakers. *International Journal of the Sociology of Language* 256: 103–128.

Stormbom, Charlotte. 2024. Epicene pronouns new and old. In Laura Louise Paterson ed., 411–420.

Straaijer, Robin. 2014. *Hyper Usage Guide of English (HUGE-database)*. Leiden: Leiden University.

Styczynski, Tomasz, Jagoda Sadlok and Jan Styczynski. 2023. Hematology on Twitter. *Acta Haematologica Polonica* 54/1: 6–10.

Tausanovitch, Chris and Christopher Warshaw. 2014. Representation in Municipal Goverment. *American Political Science Review* 108/3: 605–641.

Thelwall, Mike, Saheeda Thelwall and Ruth Fairclough. 2021. Male, female, and nonbinary differences in UK Twitter self-descriptions: A fine-grained systematic exploration. *Journal of Data and Information Science* 6/2: 1–27.

Tieken-Boon van Ostade, Ingrid. 2020. *Describing Prescriptivism. Usage Guides and Usage Problems in British and American English*. London: Routledge.

Tyrkkö, Jukka, Magnus Levin and Mikko Laitinen. *Actually* in Nordic tweets. *World Englishes* 40/4: 631–649.

Tucker, Liam and Jason J. Jones. 2023. Pronoun lists in profile bios display increased prevalence, systematic co-presence with other keywords and network tie clustering among US Twitter users 2015–2022. *Journal of Quantitative Description: Digital Media 3*. https://doi.org/10.51685/jqd.2023.003

Venkatraman, Sakshi. 2020. *Beyond 'He' and 'She': 1 in 4 LGBTQ Youths Use Nonbinary Pronouns, Survey Finds*. *NBC News*. 30 July 2020. https://www.nbcnews.com/feature/nbc-out/beyond-he-she-1-4-lgbtq-youths-use-nonbinary-pronouns-n1235204 (20 November, 2023.)

Wikipedia. 2023. *Gender Neutrality in Languages with Gendered Third Person pronouns*.https://en.wikipedia.org/wiki/Gender_neutrality_in_languages_with_gendered_third-person_pronouns#Historical,_regional,_and_proposed_gender-neutral_singular_pronouns (20 November, 2023.)

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

Zimman, Lal. 2017. Transgender language reform: Some challenges and strategies for promoting trans-affirming, gender-inclusive language. *Journal of Language and Discrimination* 1/1: 84–105.

Zimman, Lal. 2019. Trans self-identification and the language of neoliberal selfhood: Agency, power, and the limits of monologic discourse. *International Journal of the Sociology of Language* 256: 147–175.

Corresponding author
Lucía Loureiro-Porto
Universitat de les Illes Balears
Departament de Filologia Espanyola, Moderna i Clàssica
Facultat de Filosofia i Lletres
Edifici Ramon Llull
Cra. de Valldemossa Km. 7,5
Palma 07122
Spain
E-mail: lucia.loureiro@uib.es

received: November 2023
accepted: July 2024

RiCL Research in Corpus Linguistics

A dialectological approach to complement variability in global web-based English

Raquel P. Romasanta
University of Santiago de Compostela / Spain

Abstract – Computer-Mediated Communication is part of the everyday lives of a great many people of all ages, cultures, social statuses, and geographical locations. In the present study, I explore non-categorical syntactic variability in internet language with data from the *Corpus of Global Web-Based English* (GloWbE), which includes material from blogs, forums, comments, and other types of websites. The focus is on how the geographical area of internet users affects the use of the clausal complementation patterns available for the verb REGRET. The analysis of more than 10,000 examples from Indian, Sri Lankan, Pakistani, Bangladeshi, Singaporean, Malaysian, Philippine, Hong Kong, British, and American Englishes shows that geographical origin does have a bearing on the complementation system of this verb, in terms of both the factors that determine variability and the preferences for particular patterns. The varieties displaying more similarities are those that are geographically close, making the distinction between three geographical areas possible: South Asia (India, Sri Lanka, Pakistan, and Bangladesh), South-East Asia (with Singapore, Malaysia, and the Philippines) and East Asia (Hong Kong).

Keywords – computer-mediated communication; complementation; World Englishes; language contact; geographical proximity; transfer.

1. INTRODUCTION[1]

Santoro (1995: 11) defines Computer-Mediated Communication (henceforth, CMC) as encompassing all computer uses, including statistical and financial programs, remote-sensing systems, and so on, and Herring (1996: 1) defines it as "communication that takes place between human beings via the instrumentality of computers." Nowadays, when we talk about CMC, we focus mainly on the communication through and about the internet and web, including instant messaging, video conference, email, social media, and the World Wide Web. This work draws on data from the web for the study of a grammatical construction across Englishes around the world, in particular, clausal complementation

[1] I would like to express my appreciation to the two anonymous reviewers and the editors whose constructive comments improved the quality of the paper considerably. Any errors remain my sole responsibility. For support with this study, my gratitude goes to the *Spanish Ministry of Science and Innovation* (grant PID2020–117030GB–I00 funded by MCIN/AEI/10.13039/501100011033), and the *Recovery, Transformation, and Resilience Plan of the European Union NextGenerationEU* (University of Vigo, grant ref. 585507).

after the verb REGRET. Data is taken from the *Corpus of Global Web-Based English* (GloWbE; Davies and Fuchs 2015a), which includes blogs, forums, comments, and other types of websites from 20 different countries.

A previous study (Romasanta 2021) in Asian varieties on the complementation profile of REGRET, which allows non-categorial variation between finite (*that*) and nonfinite (*-ing*) complement patterns with anterior (1) and simultaneous (2) meanings, finds similar distributions of complements across varieties.

(1) a. This is when you will hugely **regret** *that* you went to Lahore to attend your second cousin's…

b. This is when you will hugely **regret** *going* to Lahore to attend your second cousins third marriage to a half Iranian-half Pakistani woman brought up in the US, because you thought it would be a lark. (GloWbE-BD)

(2) a. In these circumstances the Secretary of State **regrets** *that* he is not prepared to extend your stay to enable you to continue as a student at one of the Hubbard Colleges. (GloWbE-HK)

b. In these circumstances the Secretary of State **regrets** not *being* prepared to extend your stay to enable you to continue as a student at one of the Hubbard Colleges.

For example, Pakistani and Sri Lankan Englishes have a clear preference for finite patterns, with 57 percent and 55 percent, respectively, and Hong Kong, Bangladeshi, and Indian Englishes prefer nonfinite complements, with 59 percent, 59 percent, and 61 percent, respectively. The author hypothesizes that these similarities might be explained by the complement constructions available in the substrate languages spoken in each region since many times the effects of language contact do not surface as direct structural transfer from the indigenous languages to the target language, but rather as differences in frequencies of use and preference for some patterns over others, which makes its identification more difficult (see also Thomason 2001; Gut 2011; Brunner 2014, 2017; Romasanta 2021). Romasanta (2021: 1162) concludes that the substrate languages do not seem to explain the similarity of distributions since varieties with the same complementation systems show similar distributions to those with different systems. Other hypotheses briefly mentioned in the study without any statistical tests applied are the evolutionary development of the individual varieties and geographical proximity.

The present study focuses on the latter hypothesis, geographical proximity of English varieties. That is, on how the geographical area of internet users affects the use

of the clausal complementation patterns available for this verb, not only in terms of the distribution of the patterns but also the intra-linguistic conditioning factors affecting the speakers' choice. Regarding the geographical areas, I distinguish between South Asia (India, Sri Lanka, Pakistan, and Bangladesh), South-East Asia (Singapore, Malaysia, and the Philippines), and East Asia (Hong Kong), and the United States and Great Britain as a baseline. The aim is to test the fundamental principle of dialectology that states that "geographical proximity between dialects should predict dialectal similarity between dialects" (Szmrecsanyi 2013: 837). In other words, we can expect geographically close varieties to exhibit more similarities than distant ones, and, in principle, this should be the case for the language used on the internet, as it is elsewhere. Therefore, South Asian varieties should exhibit more similar complementation preferences when compared to the South-East Asian ones. A study of the distribution of the aforementioned finite and nonfinite complementation patterns, not only in general numbers but also in terms of the factors that influence the choice through non-hierarchical phylogenetic networks (*NeighborNet*; Bryan and Moulton 2004), will help me to test this principle.

The paper is structured as follows. Section 2 provides an overview of the extra-linguistic factors that might be at play in syntactic variability, i.e., geographical proximity, second language acquisition (henceforth SLA) processes (transparency and transfer from substrate languages), and evolutionary phase of development. Section 3 describes the data selection, annotation, and analysis. Section 4 discusses the results of the study and is followed by the conclusion in Section 5.

2. THEORETICAL BACKGROUND

English varieties around the world, or World Englishes, have been described in the literature as independent varieties of English in their own right, as opposed to simple deviations from British English (Platt *et al.* 1984), and as exhibiting similarities to other English varieties (Strevens 1980: 85).

The study of geography as a determining factor of similarities across dialects, although frequently neglected in the study of World Englishes, is a common practice in dialectology studies and one of the extra-linguistic dimensions along which English varieties are commonly aligned (Szmrecsanyi and Röthlisberger 2019; Szmrecsanyi and Grafmiller 2023). The focus of the present study is not the analysis of dialects of English

in the traditional sense; however, it seems plausible that geographical proximity might also predict similarity between varieties of English. In studies on World Englishes, this was raised as early as 1980 in Streven's *World Map of English Model*, in which he mentions that each form of English "normally exhibits similarities with other forms of English in the same geographical area" (Strevens 1980: 85). However, there has been little work that considers geography as a potential predicting factor for similarities and dissimilarities across global varieties of English. Of the very few authors who have done so, Szmrecsanyi and Kortmann (2009b) and Szmrecsanyi (2013) find geography to be a weak predictor of variability. Szmrecsanyi (2013: 841), for example, in a study of morphosyntactic similarities in L1 varieties, finds that geography accounts for less than five percent of the variability found and that there is a typological split "between traditional L1 varieties, high-contact L1 varieties, and what we have dubbed 'higher-contact' L1 varieties of English (such as the AAVE varieties)." In contrast, Kortmann and Schröter's (2017: 308) *NeighborNet* analysis of the survey data from the *World Atlas of Variation in English* project yields evidence of regional clustering, for example, South Asian and South-East Asian varieties in the same cluster but in different branches. In this direction, Fuchs *et al.* (2019) look at the present perfect in African English varieties, British, American, and Philippine English, and also find geographical proximity as the most important predictor.[2]

In the remainder of this section, I will briefly describe the other extra-linguistic factors that might affect the English varieties around the world previously mentioned.

2.1. SLA and language contact processes

There are two main processes that I would like to discuss here: 1) the principle of maximization of transparency and 2) language transfer. The principle of maximization of transparency is one of the production principles mentioned by Williams (1987).[3] Slobin (1980) considers transparency as the one-to-one mapping of form and meaning, that is, an intended underlying meaning is expressed with one clear, "invariant surface form (or

[2] As a reviewer rightly pointed out, the highly active work on epicenter theory in World Englishes relates to this argument. However, as the present study focuses on language-use data, it will not be possible to identify the influence of a variety on another. In order to do so, a mixed-method approach, including attitudinal data as well as historical background data, is necessary (Hundt 2013: 184; Peters and Bernaisch 2022).

[3] Also referred to as the 'one-to-one principle' in Andersen (1984), 'iconicity' in Haiman (1985), and 'isomorphism 'in Givón (1985).

construction)" (Andersen 1984: 79). World Englishes are said to show a tendency towards transparency because transparent constructions are easier for the speaker to produce and for the listener to parse (Slobin 1973, 1977; Karmiloff-Smith 1979; Williams 1987: 179). Multiple studies have focused on this tendency for transparency (see, for example, Williams 1987; Szmrecsanyi and Kortmann 2009a; Steger 2012; Romasanta 2017). In the complementation system, this was attested within the alternation between finite and nonfinite clauses. Finite complement clauses are more transparent because they are marked for tense, agreement and modality, have an explicit subject, and usually a complementizer, and therefore the relationship between form and meaning is tighter than in nonfinite clauses (Givón 1985: 200; Schneider 2012a, 2013; Steger and Schneider 2012; Romasanta 2017, 2019). Therefore, in the present study, I will test this tendency for transparency by looking at the distribution between finite and nonfinite patterns with the verb REGRET.

The other SLA and language contact process —and probably the most obvious contact-induced change— is transfer. At the level of grammar, Schneider (2007: 83) argues that innovations occur mainly at the interface between lexis and grammar, a classic example being verb and adjective complementation, and indeed a series of studies have focused on the innovations present in the complementation system (see Mukherjee and Hoffmann 2006; Mufwene and Gries 2009; Deshors and Gries 2016; Gries and Bernaisch 2016, among others). In order to find this language transfer, we must know the complementation systems of the different substrate languages spoken in each region. Methodologically, this brings up some difficulties. Firstly, it is impossible to assign a particular substrate language to a particular speaker in the GloWbE data, and, secondly, the number of substrate languages in some countries goes beyond the hundreds, so I could only look at the ones with a written tradition. This means that conclusions for the effect of language transfer must be taken with care. In what follows, I will briefly describe the complementation systems of the main substrate languages in each region, although we must not forget that the sociolinguistic situations of these regions are more complex than can be described in detail here (see Table 1 in Section 2.2 for a summary of the substrate languages and the phases of development of each variety).

Based on the *World Factbook* (CIA 2024), the dominant substrate languages in **India** are Hindi, Bengali, Marathi, and Telugu. Even though many other languages are also part of the sociolinguistic landscape —for example, Tamil, Gujarati, Urdu, Kannada,

Malayalam, Punjabi, among others— I will focus on the first four since they are the most widely spoken languages. All four languages (Hindi, Bengali, Marathi, and Telugu) use finite clauses in their complementation system, and these are marked with the complementizers *ki*, *bôle*, *ki*, and *ani*, respectively. Three of these languages (Hindi, Marathi, and Telugu) have nonfinite complements, which consist of the suffixes *na:*- in Hindi, -*aTam* in Telugu, and -*āy*,-*ūn*, and -*lyā* in Marathi, added to the verb stem (see Koul 2008: 181–185 for Hindi, Krishnamurti and Gwynn 1985: 234, 363 for Telugu, Pandharipande 1997: 65–68, 444 for Marathi).

In **Sri Lanka**, the main substrate languages are Sinhala and Tamil (CIA 2024). In Sinhala, finite complement clauses can be constructed with the complemetizer *kiəla*, and nonfinite complements with the complementizers *bawə* and *ekə* with a nonfinite verb (Wheeler *et al.* 2005: 173–174). In Tamil, finite complements take the complementizer *nuu* (Schiffman 1999: 152, 174), and nonfinites are constructed adding the suffixes -*a*, -*tu*, -*ntu*, -*ttu*, or -*i* to the verb stem (Lehmann 1993: 71–72).

According to the *World Factbook* (CIA 2024), the dominant substrate languages in **Pakistan**, are Punjabi, Pashto, and Sindhi. Finite complements are constructed with the marker *ki* in Punjabi, *tse* or *che* in Pashto, and *ta* in Sindhi. The suffixes -*Naa*/*naa* and -*an.u*/*in.u*, in Punjabi and Sindhi, respectively, are used for nonfinite complementation (see Bhatia 1993: 44, 50 for Punjabi, Tegey and Robson 1996: 199 for Pashto, and Yegorova 1971: 74–75 for Sindhi).

Bengali, also known as Bangla, is the dominant substrate language in **Bangladesh** (CIA 2024). As mentioned previously, Bengali has only finite complements.

In **Singapore**, the dominant substrates are Mandarin and other Chinese dialects (including Hokkien, Cantonese, Teochew, Hakka; CIA 2024). These languages use the juxtaposition of clauses, so neither finite nor nonfinite complementation is possible here (see, for example, Haspelmath *et al.* 2001: 979 for Mandarin, Fang 2010: 104 for Hokkien, and Matthews and Yip 1994: 174, 293 for Cantonese).

In **Malaysia**, the main substrate languages are Malay and a number of Chinese dialects. As mentioned previously, the Chinese dialects do not have finite or nonfinite complementation. In Malay, finite complement clauses take the complementizer *bahawa* (Omar and Subbiah 1989: 97) while nonfinite clauses do not exist (Nordhoff 2009: 276–279).

According to the *World Factbook* (CIA 2024), the dominant substrate language in the **Philippines** is Tagalog, where finite clauses are introduced by the linker *na*/*-ng* (Schachter and Otanes 1972: 172).

Cantonese is the dominant substrate language in **Hong Kong**, 88.9 percent, together with Mandarin and other Chinese dialects (CIA 2024). As already stated, these Chinese dialects do not have finite or nonfinite complementation.

2.2. Evolutionary phase of development in the Dynamic Model (Schneider 2007)

The most widely discussed model of classification of World Englishes is the 'Dynamic Model' (Schneider 2007).[4] The main assumption here is that the different post-colonial Englishes undergo the same uniform process of identity reconstruction divided into five phases: foundation, exonormative stabilization, nativization, endonormative stabilization, and differentiation (Schneider 2007: 30–35). Various earlier studies found a correlation between phase of development in this model and degree of complexity. Research on verb complementation in particular shows mixed results regarding this correlation. Mukherjee and Gries (2009: 48–49) study ditransitive, monotransitive, and intransitive constructions in Hong Kong, Indian, and Singaporean English showing that the correlation holds true: "the more advanced a New English variety is in its evolution, the more dissimilar it is to British English at the level of collostructions." Schneider (2012b) looks at the alternation between finite and nonfinite clauses with several number of verbs, taking into account the presence or absence of the complementizer *that* and an explicit modal. His results also confirm the correlation in that they indicate that less advanced varieties, in this case Hong Kong English and East African English, have a stronger tendency to use simpler patterns than the more advanced ones, Singaporean and Indian Englishes. However, the correlation is not found in Deshors and Gries' (2016) study of *-ing* and *to*-infinitive complement alternation in Singaporean, Hong Kong, and Malaysian Englishes. The most advanced variety, Singaporean English, is not dissimilar, but in fact the most similar to the native Englishes (British and American English). In a similar vein, García-Castro (2018) and Romasanta (2019, 2021) study complement variability with the retrospective verbs REMEMBER and REGRET, respectively, and also detect stronger preferences for

[4] Other models of classification frequently alluded to are also available. For example, Kachru (1982), Mair (2013) and, more recently, Buschfeld and Kautzsch (2017).

simpler finite clauses in less advanced varieties and for more complex nonfinite patterns in the more advanced varieties. The greater use of nonfinite patterns in more advanced varieties, therefore, makes them more similar to British English.

It seems suitable then to briefly consider the evolutionary phase of development in the Dynamic Model of each Asian variety included in the study to assess the potential effect on the alternation between finite and nonfinite complementation. Table 1 below summarizes this. Two important notes are in point. Firstly, Singapore is in phase 4 in the Dynamic Model, endonormative stabilization. However, it is said to have become a first/native language (L1), with many of its young speakers learning it as their first language, so that it is gradually developing from ESL to ENL (Gupta 1994; Lim and Foley 2004; Tan 2014; Lim 2017; Buschfeld 2020a, 2020b). Secondly, regarding Hong Kong, Schneider (2007: 133) claims that it has "reached stage 3 [but] with some traces of phase 2 still observable," and Setter *et al.* (2010: 116) argue that "Hong Kong English will eventually be pushed more firmly towards Kachru's Outer Circle, Schneider's phase 4." Until the handover of the territory to China in 1997, English was the medium of instruction in most schools, but a change in policy then ensued. There has since been a process of mainlandization by which the government has begun to favor the use of Cantonese as the medium of instruction, while reducing the number of schools allowed to use English.

		Complementation			**Evolutionary**
	Variety	**Finite**	**Nonfinite**	**Summary**	**phase**
South Asia	India	Yes	Yes	Both	3+
	Sri Lanka	Yes	Yes	Both	4
	Pakistan	Yes	Yes	Both	3+
	Bangladesh	Yes	No	Finite	2+
South-East Asia	Singapore	No	No	None	4
	Malaysia	Yes	No	Finite	3
	The Philippines	Yes	?	Finite	4
East Asia	Hong Kong	No	No	None	3

Table 1: Summary of the substrate languages and the phase of development of each Asian variety of English

3. DATA AND METHODOLOGY

3.1. The corpus

The data has been taken from the GloWbE corpus (Davies and Fuchs 2015a), an online corpus released in 2015 with 1.9 billion words from 1.8 million web pages in 20 different countries (United States, Canada, Great Britain, Ireland, Australia, New Zealand, India, Sri Lanka, Pakistan, Bangladesh, Singapore, Malaysia, Philippines, Hong Kong, South Africa, Nigeria, Ghana, Kenya, Tanzania, and Jamaica).

In order to identify the countries of origin of each web page, they carried out the searches for each country separately relying on *Google's Advance Search*, which relies on country domains as well as on "the IP address for the web server, who links to that website, and who visits the website" (Davies and Fuchs 2015b: 4). This, however, has been criticized several times since country domains such as *.to* (Tonga) may retrieve websites from Tokyo, Toronto, or Timbuctoo, as well as websites such as www.knowhow.to or www.invitation.to. Even if the website is correctly cataloged, the writer may not be originally from the country (Nelson 2015: 39; Deshors and Bernaisch 2019). This also has an impact on the researchers' knowledge of the writer's backgrounds (age, gender, mother tongue, etc.), which is especially relevant for the present study as one of the hypotheses is related to the substrate languages of the writers. From a methodological perspective, the study of the substrate languages poses a problem, and, therefore, conclusions on this matter are to be taken with care.

Despite of the issues mentioned above, I see the GloWbE corpus as "a big and aggregative corpus" (Brezina and Meyerhoff 2014; Mukherjee 2015: 36) and expect that its size will statistically overcome its hindrances (Davies 2012; Nelson 2015: 39; Hundt 2020). In fact, studies based on GloWbE that replicate earlier studies carried out with smaller corpora obtain similar results (see, for example, Heller and Röthlisberger 2015).

3.2. Manual data pruning and coding

For this study, data represents eight different English varieties from the Asian continent, namely Indian, Sri Lankan, Pakistani, Bangladeshi, Singaporean, Malaysian, Philippine, and Hong Kong Englishes, and the two main metropolitan varieties, British and American English in the GloWbE corpus (regret*_v*). The total number of examples retrieved was 10,275.

After the manual pruning of the examples, I codified all relevant instances according to 11 intra-linguistic conditioning factors. The list is as follows:

1. Meaning of the verb in the MC (main clause) (dichotomous: regret1, regret2).
2. Meaning of the verb in the CC (complement clause) (dichotomous: action, state).
3. Animacy of the subject in the CC (dichotomous: animate, inanimate).
4. Type of subject in the MC (qualitative: pron1, pron2, pron3, NP, none).
5. Type of subject in the CC (dichotomous: complex noun phrase (CNP), other).
6. Voice of the CC (qualitative: active, passive, copular).
7. Polarity of the CC (dichotomous: positive, negative).
8. Complexity of the CC (quantitative: number of words).
9. Presence of intervening material (quantitative: number. of intervening words).
10. Subject coreferentiality (dichotomous: coreferential, non-coreferential).
11. Horror aequi (dichotomous: *yes*, no).

The first three of these are semantic factors. The two meanings of the verb in the MC are taken from Cuyckens *et al.* (2014: 188) where they define 'regret1' as "to feel sorry about something one has done and that one should have done differently or about a state of affairs one is involved in or responsible for and that one wishes was different", as in (3), and 'regret2' as a "a more 'polite' use of REGRET where the speaker says that s/he is sorry or sad about a situation, usually one that s/he is not directly responsible for," as in (4). For the meaning of the verb in the CC, the distinction between action and state was drawn from Quirk *et al.* (1985: 201), see examples (5) and (6), respectively. Lastly, for the animacy of the subject in the CC, I used a binary classification distinguishing between animate (7) and inanimate (8).

(3) Tepco conference starting now: "We **regret** that we are causing concern to many residents of Japan." (GloWbE-US)

(4) We **regret** that our client was not provided with more time. (GloWbE-LK)

(5) One thing I know is that I never **regret** *attending* this course. (GloWbE-MY)

(6) He **regrets** not *having* the chance to tour the Philippines yet, things that made him feel… (GloWbE-PH)

(7) We do **regret** *that the terrorists* were actually horrific acts and they were terrorist acts. (GloWbE-PK)

(8) Bradley has since publicly stated he was humbled by the Morton case and **regrets** *his actions* opposing DNA testing in the case. (GloWbE-US)

The next seven factors (from four to ten above) are features relating to processing complexity. These are important for the alternation, since with complement clauses involving higher processing complexity, speakers generally prefer more grammatically explicit constructional variants ('Complexity Principle'; Rohdenburg 1996). The complement clause can be active, passive, or copular, as in (9), (10), and (11), respectively, and negative (12) or positive. Then, I coded the complexity of the complement clause (13), which contains a total of 87 words, and the presence of intervening material (14), which has six words as intervening material. In terms of subject coreferentiality between the main and complement clauses, these can be coreferential (15) or non-coreferential (16). Finally, the last factor exemplifies the generalization known as the 'Horror Aequi Principle', which holds that speakers tend to avoid (near-)identical and (near-)adjacent structures (Brugmann 1909; Rohdenburg 2003). This factor has two levels, 'yes' (17), when there is an environment where this principle might be at work, and 'no'.

(9) I **regret** that I have *wasted* about 2 weeks on this site trying to reason and arrive at some kind of consensus which would move Sri Lanka forward, … (GloWbE-LK)

(10) … that they might have no cause to **regret** *being denied* the option of any other. (GloWbE-GB)

(11) Have you ever **regretted** *being* a monk? (GloWbE-MY)

(12) The meeting **regretted** that India was *not* interested in the resumption of dialogue. (GloWbE-PK)

(13) I **regret** *that the U.S. has suffered itself to be brought so low by the vultures and crooks who are operating the roulette wheels and faro tables in the Fed, that is now obliged to throw itself on the mercy of its legislators and charwomen, its clerks, and it poor pensioners and to take money out of our pockets to make good the defalcations of the International Bankers who were placed in control of the Treasury and given the monopoly of U.S. Currency by the misbegotten Fed*. (GloWbE-US)

(14) I **regret**, *from a personal point of view*, being here. (GloWbE-GB)

(15) *She* constantly **regrets** that *she* could not afford to send her daughters to school during the hard times, … (GloWbE-BD)

(16) …although *we* **regret** *her* not coming to Asia-Pacific, so that she could address this… (GloWbE-LK)

(17) I am **regretting** *writing* it but can't stop because you deserve it. (GloWbE-IN)

3.3. Statistical analysis

Data was subjected to non-hierarchical phylogenetic networks (*NeighborNet*) as an exploratory method to visually represent which varieties are more similar and whether this could correspond to geographical proximity. This is a clustering method originating in bioinformatics (Bryant and Moulton 2004) and frequently used in historical, dialectological, and typological linguistics (McMahon and McMahon 2005; Cysouw 2007; McMahon *et al.* 2007; Szmrecsanyi and Wolk 2011). These networks allow for a more fine-grained analysis, as compared to other multidimensional aggregation analyses such as hierarchical cluster analysis, as they "produce an unrooted network representation (NeighborNet) that establishes, first of all, "geolinguistic signal[s]" (Szmrecsanyi 2013) in the data" (Werner 2014).

The analysis was conducted in *R* (R Core Team 2022) using the *NeighborNet* package (Ansari and Draghici 2019). These have been shown to be a great tool to graphically represent relationships of similarity and dissimilarity between multiple objects. Each object, here English varieties, represents its own cluster. They are compared pairwise within a distance matrix and the most similar ones are merged until all objects are merged into one tree. In order to create the distance matrix, I used the relative values of the individual factors and, to measure distances and similarities between varieties of English, I used the Euclidean distance, which in the case of the present dataset is fully proportional to the Manhattan distance. The Euclidean distance measure "is similar to our everyday idea of the distance between two objects", where we would take the shorter direct route (see Figure 1 below; Levshina 2015: 306–307). The resulting networks are unrooted family trees so that the length of each branch is proportional to linguistic distances (Bryant and Moulton 2004; Szmrecsanyi 2013: 841). This means that proximity in the net indicates similarity in the complementation profile of REGRET in the varieties of English.

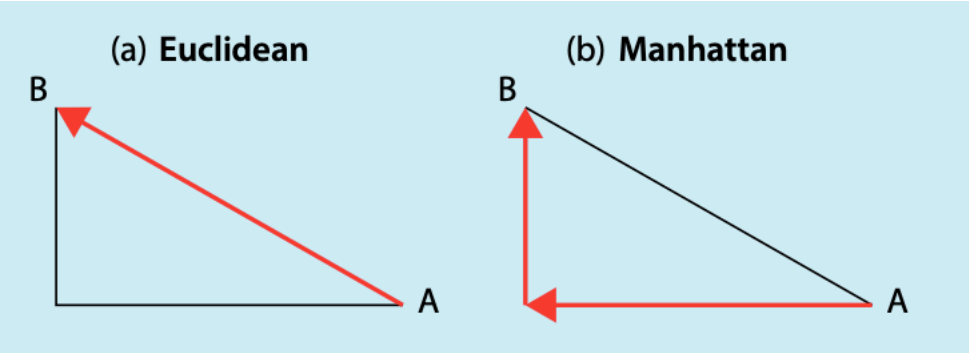

Figure 1: Distance metrics: a) Euclidean, b) Manhattan (from Levshina 2015: 307)

4. Results and discussion

I begin the data analysis with an overview of the distribution of finite *that*-clauses and nonfinite *-ing* clauses across varieties. As can be seen in Figure 2, British (GB) and American English (US) have the same distribution, with a clear preference for nonfinite clauses (73%) over finite ones (27%). The next three varieties, Singaporean (SG), Malaysian (MY), and Philippine Englishes (PH) have a very similar distribution to the metropolitan varieties, or even the same as in the case of the Philippines. Compared to British and American Englishes, Singaporean English shows a slightly stronger preference for *-ing* clauses, 78 percent. Then, Malaysian and Philippine Englishes have a very similar distribution, with 74 percent and 73 percent of nonfinite clauses, respectively. The remaining varieties, Indian (IN), Bangladeshi (BD), Hong Kong (HK), Sri Lankan (LK), and Pakistani Englishes (PK) show a stronger use of *that*-clauses, with 39 percent, 41 percent, 41 percent, 54 percent, and 57 percent, respectively, as compared to British and American Englishes, which only have 27 percent of finite clauses.

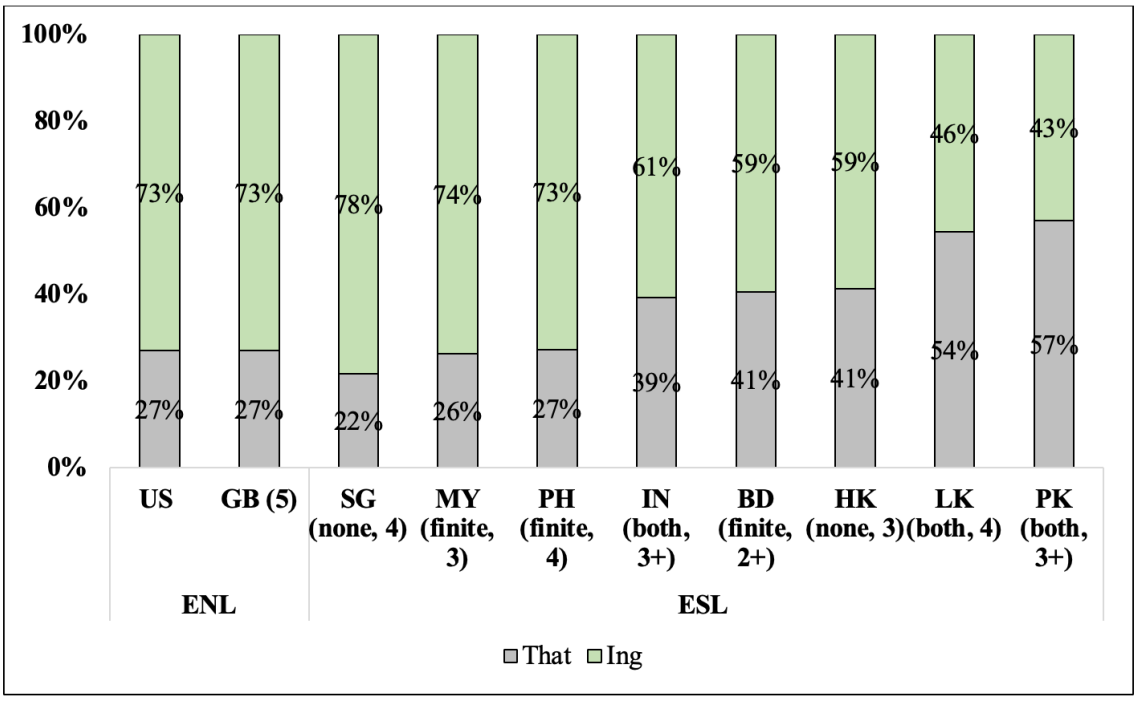

Figure 2: Distribution of finite *that*-clauses and nonfinite *-ing* clauses across Asian varieties of English

The greater use of finite patterns in India, Bangladesh, Hong Kong, Sri Lanka, and Pakistan might be the effect of the SLA strategy of maximization of transparency described in Section 2.1. According to this, ESL speakers would prefer transparent constructions due to these being easier to produce and parse. However, looking at the data in Figure 2, this tendency towards transparency would not explain why there is a stronger preference for *that*-clauses in some varieties. The explanation does not seem to lie on the transfer effect from substrate languages since, as can be seen in brackets, varieties with similar complementation systems in their substrate languages have different distributions between finite and nonfinite patterns. See, for example, the distributions in Singapore and Hong Kong. Complementation in the substrate languages in both regions is constructed through parataxis, that is, the juxtaposition of two clauses so that "the two clauses are more symmetrical than main and subordinate clauses in English" (Matthews and Yip 1994: 293). However, Singapore shows a clear preference for the use of *-ing* clauses (78%) while this preference is reduced to 59 percent in Hong Kong. The same occurs with India and Sri Lanka or Pakistan, where finite and nonfinite complement constructions are available in the substrate languages. While, in India, there is a stronger use of *-ing* clauses (61%), in Sri Lanka and Pakistan, the preference is for *that*-clauses (54% and 57%, respectively).

Another potential explanatory extra-linguistic factor mentioned in Section 2.2. is the effect of the evolutionary phase of development in terms of Schneider's Dynamic Model (2007). According to different studies, there should be a stronger preference for simpler *that*-clauses in less advanced varieties and for more complex *-ing* clauses in the more advanced varieties (Schneider 2012b; Brunner 2017; García-Castro 2018; Romasanta 2021). Looking back at Figure 2, less advanced varieties such as Hong Kong and Bangladesh, in phases 3 and 2+, have a stronger preference for simpler *that*-clauses (41%), as compared to Great Britain, with 27 percent. However, Malaysia, which is another variety in phase 3, shows a clear preference for more complex *-ing* clauses (74%). The more advanced varieties, Singapore and Philippines, both in phase 4, have a clear preference for the use of complex *-ing* clauses (78% and 73%, respectively), but other varieties, such as Sri Lanka and Pakistan, in phases 4 and 3+, prefer *that*-clauses, 54 percent and 57 percent, respectively. Therefore, the evolutionary phase of development does not seem to fully account for the different distributions of finite and nonfinite complement patterns across English varieties.

Figure 3 below is the output of the non-hierarchical phylogenetic network (*NeighborNet*) where distances between varieties are represented considering the distribution of finite and nonfinite patterns and the 11 intra-linguistic conditioning factors described in Section 3.1. Each node is one English variety, here referred to by their respective abbreviations, and information regarding the phrase of development and the presence or absence of finite and/or nonfinite clauses in the dominant substrate languages in the parentheses. When there are finite and nonfinite complements in the substrates, I use 'both'. 'Finite' is used when only finite clauses are possible, and 'none' when complementation is constructed through other strategies and neither finite nor nonfinite complements exist. The diagram is self-explanatory and can be basically read like a family tree that is not rooted; branch lengths are proportional to linguistic distances. A long path therefore indicates many differences, while a short path indicates that the varieties are fairly similar. Sets of parallel lines and boxy shapes indicate splits in the data. Starting with the top section of Figure 3, we find the Philippines, Singapore, and Malaysia, together with the two metropolitan varieties, Great Britain and the United States. From this group, it is necessary to highlight the connection between the Philippines and the United States, since the Philippines is the only American colony included in this study. We should also point out Singapore within this group, since some signs of it becoming an L1 are visible (Buschfeld 2020a), which, together with the trend towards

the americanization of English (Buschfeld and Kautzsch 2017; Gilquin 2018; Gonçalves *et al.* 2018; Low and Pakir 2018), might explain its proximity to the United States. In terms of the substrate languages of this group, this figure confirms what was already discussed with Figure 2, that is, transfer of features from the substrate languages spoken in each region does not seem to be an explanatory factor. As can be seen in brackets next to each variety, the Philippines and Malaysia, both with finite complements in their substrates, are located near Singapore, which does not have clausal complementation. The varieties in this upper section of the figure are also in a mixture of phases of the Dynamic Model (Schneider 2007); Malaysia in phase 3, the Philippines and Singapore in phase 4, and the United States in phase 5. This also confirms that the evolutionary phase of development does not seem to explain the closeness of the varieties in the figure. On the other hand, if we look at the varieties in this group in terms of their geographical location, Singapore, Malaysia, and the Philippines are in what is commonly referred to as South-East Asia. Therefore, it seems that their geographical proximity may be behind their similarities.

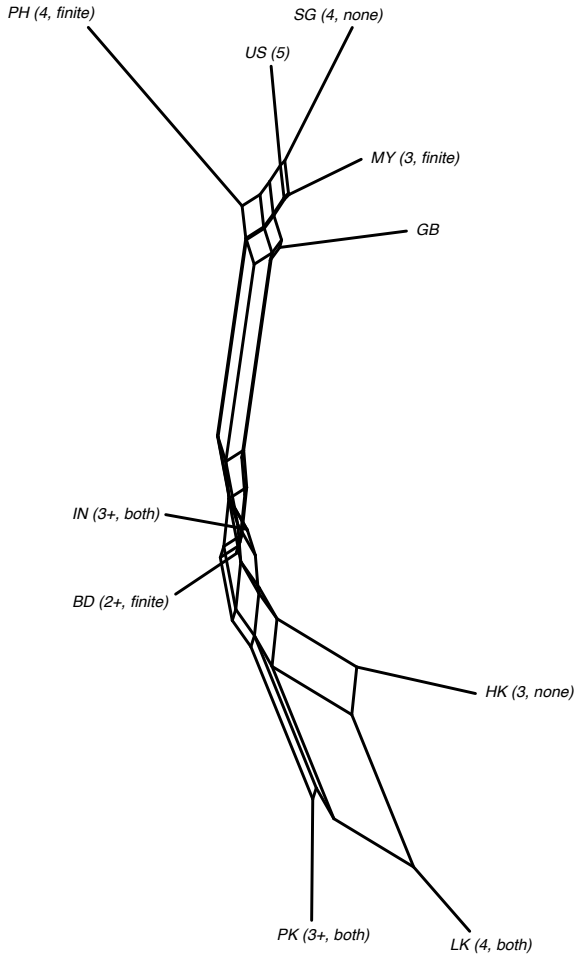

Figure 3: *NeighborNet* of similarity across Asian varieties of English

A look at the bottom section of the figure shows a similar picture. In this section, we find India, Bangladesh, Pakistan, Sri Lanka, and Hong Kong. First, it is important to highlight that there is an important historical connection between India, Pakistan, and Bangladesh that cannot be ignored; during the British Empire and, therefore, when English was introduced in the region, these three countries were one nation. However, regarding the Dynamic Model (Schneider 2007), these are in different phases. India and Pakistan are in an advanced stage of phase 3, while Bangladesh is still in phase 2. Additionally, in this group we also have Hong Kong in phase 3, and Sri Lanka in phase 4. Therefore, here the phase of development seems not to be sufficient to explain similarities and differences between varieties. As for the substrate languages, they do not seem to explain the proximity of the varieties since, in this group, there are English varieties with both finite and nonfinite complementation systems in their substrate languages (India, Pakistan, and Sri Lanka), one with only finite complements (Bangladesh), and one with no clausal complementation (Hong Kong). What does seem to explain the closeness between varieties, and therefore their similarities, is the geographical location. India, Bangladesh, Pakistan, and Sri Lanka are South Asian varieties, and Hong Kong, a little further away in the figure, is part of East Asia.

Therefore, from Figures 2 and 3, it can be concluded that the phase of development in Schneider's Dynamic Model (2007) and the transfer of features from the substrate languages —the major factors frequently studied in the literature as determinants of the variation in World Englishes— do not seem to account for the similarities and differences between the varieties of English studied here. If we look at the varieties individually, it may seem that these extra-linguistic factors could explain the preference for less complex structures within a non-categorical variation in ESL varieties. However, when studying a greater number of English varieties, it can be noticed that varieties in different evolutionary phases of development and with different complementation systems are similar in terms of their choice of less complex structures, which demonstrates that, at least in this case, these two factors are not as decisive as they may seem at first glance. On the contrary, a factor such as geographical location, which has not been studied very often and that cannot be taken into account with investigations of individual varieties, does seem to have a greater explanatory power of the similarities across English varieties.

5. CONCLUSION

This paper is a step forward in the study of CMC by analyzing the English used on the internet. The study analyzed more than 10,000 examples of the complementation of the verb REGRET on the GloWbE corpus in Asian varieties of English (India, Sri Lanka, Pakistan, Bangladesh, Singapore, Malaysia, the Philippines, and Hong Kong) and metropolitan varieties (Great Britain and the United States). There was a special focus on geographical proximity of the varieties as a potential extra-linguistic determining factor for the similarities and differences found, even though other factors frequently discussed in the literature —such as SLA and language contact processes, and the effect of the phase of evolution of the individual varieties in terms of Schneider's Dynamic Model (2007)— were also considered.

The non-categorical variability with this verb is between finite *that*-clauses and nonfinite *-ing* clauses (*you will **regret** that you went to Lahore* vs. *you will **regret** going to Lahore*). Results showed a clear different distribution of these two patterns across World Englishes, with a general preference for *that*-clauses in ESL varieties, more specifically in India, Bangladesh, Hong Kong, Sri Lankan, and Pakistani Englishes. However, there are other three varieties in which a more frequent use of *-ing* clauses can be seen, in particular, Singapore, Malaysia, and the Philippines. The principle of maximization of transparency and the transfer of features from substrate languages, the extra-linguistic factors result of the SLA process, and the phase of development in Schneider's Dynamic Model (2007), do not account for the similarities and differences between the varieties of English studied here. The non-hierarchical phylogenetic network (*NeighborNet*) has brought light to another extra-linguistic factor that has not often been studied in this area of linguistics, the geographical proximity of the varieties under research. The varieties displaying more similarities are those that are geographically close making the distinction between three geographical areas possible: South Asia (with India, Sri Lanka, Pakistan, and Bangladesh), South-East Asia (with Singapore, Malaysia, and the Philippines), and East Asia (Hong Kong).

The relevance of this study is that it has revealed the importance of the geographical location as a determining factor in the similarities and differences across World Englishes. The literature is not conclusive regarding this factor since there are studies that find it to be a weak predictor (Szmrecsanyi and Kortmann 2009b; Szmrecsanyi 2013) while in others it is the most important one (Kortmann and Schröter 2017; Fuchs *et al.* 2019). The

present investigation is another study that highlights the geographical proximity as the most important predictor. This study has also revealed the need to study larger sets of English varieties so that factors such as geographical proximity can be tested.

Future work should include a large variety of verbs so that the effect of the geographical proximity can be tested in the complementation system in general as well as focus on other linguistic features.

References

Andersen, Roger W. 1984. The one-to-one mapping principle of interlanguage construction. *Language Teaching* 34: 77–95.

Ansari, Sahar and Sorin Draghici. 2019. *NeighborNet: Neighbor_net Analysis*. *R Package* version 1.2.0.

Bhatia, Tej K. 1993. *Punjabi: A Cognitive-Descriptive Grammar*. London: Routledge.

Brezina, Vaclav and Miriam Meyerhoff. 2014. Significant or random? A critical review of sociolinguistic generalizations based on large corpora. *International Journal of Corpus Linguistics* 19: 1–28.

Brugmann, Karl. 1909. Das Wesen der lautlichen Dissimilationen. *Abhandlungen der philologischhistorischen Klasse der königlich-sächsischen Gessellschaft der Wissenschaften* 27: 141–178.

Brunner, Thomas. 2014. Structural nativization, typology and complexity: Noun phrase structures in British, Kenyan and Singaporean English. *English Language and Linguistics* 18/1: 23–48.

Brunner, Thomas. 2017. *Simplicity and Typological Effects in the Emergence of New Englishes: The Noun Phrase in Singaporean and Kenyan English*. Berlin: Mouton de Gruyter.

Bryant, Davis and Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution* 21/2: 255–265.

Buschfeld, Sarah. 2020a. *Children's English in Singapore: Acquisition, Properties, and Use*. London: Routledge.

Buschfeld, Sarah. 2020b. Language acquisition and World Englishes. In Daniel Schreier, Marianne Hundt and Edgar W. Schneider eds., 559–584.

Buschfeld, Sarah and Alexander Kautzsch. 2017. Towards an integrated approach to postcolonial and non-postcolonial Englishes. *World Englishes* 36/1: 104–126.

CIA. 2024. CIA: *The World Factbook*. https://www.cia.gov/the-world-factbook/countries/ (22 March, 2024.)

Cuyckens, Hubert, Frauke D'hoedt and Benedikt Szmrecsanyi. 2014. Variability in verb complementation in Late Modern English: Finite vs. non-finite patterns. In Marianne Hund ed. *Late Modern English Syntax*. Cambridge: Cambridge University Press, 182–204.

Cysouw, Michael. 2007. New approaches to cluster analysis of typological indices. In Peter Grzybek and Reinhard Köhler eds. *Exact Methods in the Study of Language and Text*. Berlin: Mouton de Gruyter, 61–76.

Davies, Mark. 2012. Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In Terttu Nevalainen and Elizabeth C.

Traugott eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 157–174.
Davies, Mark and Robert Fuchs. 2015a. Expanding horizons in the study of World Englishes with the 1.9-billion-word Global Web-based English Corpus (GloWbE). *English World-Wide* 36/1: 1–28.
Davies, Mark and Robert Fuchs. 2015b. A reply. *English World-Wide* 36/1: 45–47.
Deshors, Sandra C. and Stefan T. Gries. 2016. Profiling verb complementation constructions across New Englishes. *International Journal of Corpus Linguistics* 21/2: 192–218.
Deshors, Sandra C. and Tobias Bernaisch. 2019. Corpus approaches to World Englishes: A bird's-eye view. In Peter I. De Costa, Dustin Crowther and Jeffrey Maloney eds. *Investigating World Englishes: Research Methodology and Practical Implications*. London: Routledge, 21–43.
Fang, Meili. 2010. *Spoken Hokkien*. London: University of London.
Fuchs, Robert, Bertus van Rooy and Ulrike Gut. 2019. Corpus-based research on English in Africa: A practical introduction. In Alexandra U. Esimaje, Ulrike Gut and Bassey E. Antia eds. *Corpus Linguistics and African Englishes*. Amsterdam: John Benjamins, 37–69.
García-Castro, Laura. 2018. *The Complementation Profile of REMEMBER in Post-colonial Englishes*. Vigo: University of Vigo dissertation.
Gilquin, Gaëtanelle. 2018. American and/or British influence on L2 Englishes – Does context tip the scale(s)? In Sandra C. Deshors ed. *Modeling World Englishes: Assessing the Interplay of Emancipation and Globalization of ESL Varieties*. Amsterdam: John Benjamins, 187–216.
Givón, Thomas. 1985. Iconicity, isomorphism and non-arbitrary coding in syntax. In John Haiman ed. *Iconicity in Syntax*. Amsterdam: John Benjamins, 187–219.
Gonçalves, Bruno, Lucía Loureiro-Porto, José J. Ramasco and David Sánchez. 2018. Mapping the americanization of English in space and time. *PLoS ONE* 13/5: 1–15.
Gries, Stefan T. and Tobias Bernaisch. 2016. Exploring epicentres empirically: Focus on South Asian Englishes. *English World-Wide* 37/1: 1–25.
Gupta, Anthea F. 1994. *The Step-Tongue. Children's English in Singapore*. Clevedon: Multilingual Matters.
Gut, Ulrike. 2011. Studying structural innovations in New English varieties. In Joybrato Mukherjee and Marianne Hundt eds. *Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*. Amsterdam: John Benjamins, 166–205.
Haiman, John. 1985. *Natural Syntax*. Cambridge: Cambridge University Press.
Haspelmath, Martin, König Ekkehard, Wulf Oesterreicher and Wolfgang Raible. 2001. *Language Typology and Language Universals: An International Handbook*. Berlin: Walter de Gruyter.
Heller, Benedikt and Melanie Röthlisberger. 2015. Big data on trial: Researching syntactic alternations in GloWbE and ICE. Paper presented at the *Data to Evidence (d2e) Conference*. Helsinki: University of Helsinki, 21 October 2015.
Herring, Susan C. 1996. *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*. Amsterdam: John Benjamins.
Hundt, Marianne. 2013. The diversification of English: Old, new and emerging epicenters. In Daniel Schreier and Marianne Hundt eds. *English as a Contact Language*. Cambridge: Cambridge University Press, 182–203.
Hundt, Marianne. 2020. Corpus-based approaches to World Englishes. In Daniel Schreier, Marianne Hundt and Edgar W. Schneider eds., 506–533.

Kachru, Brag B. ed. 1982. *The Other Tongue: English across Cultures*. Urbana: University of Illinois Press.
Karmiloff-Smith, Annette. 1979. *A Functional Approach to Child Language*. Cambridge: Cambridge University Press.
Kortmann, Bernd and Verena Schröter. 2017. Varieties of English. In Raymond Hickey ed. *The Cambridge Handbook of Areal Linguistics*. Cambridge: Cambridge University Press, 304–330.
Koul, Omkar N. 2008. *Modern Hindi Grammar*. Springfield: Dunwoody Press.
Krishnamurti, Bhadriraju and John Peter Lucius Gwynn. 1985. *A Grammar of Modern Telugu*. New Delhi: Oxford University Press.
Lehmann, Thomas. 1993. *A Grammar of Modern Tamil*. Pondicherry: Pondicherry Institute of Linguistics and Culture.
Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.
Lim, Lisa. 2017. Southeast Asia. In Markku Filppula, Juhani Klemola and Sharma Devyani eds. *The Oxford Handbook of World Englishes*. Oxford: Oxford University Press, 448–471.
Lim, Lisa and Joseph A. Foley. 2004. English in Singapore and Singapore English: Background and methodology. In Lisa Lim ed. *Singapore English: A Grammatical Description*. Amsterdam: John Benjamins, 1–18.
Low, Ee-Ling and Anne Pakir. 2018. *World Englishes: Rethinking Paradigms*. London: Routledge.
Mair, Christian. 2013. The World System of Englishes: Accounting for the transnational importance of mobile and mediated vernaculars. *English World-Wide* 34/3: 253–278.
Matthews, Stephen and Virginia Yip. 1994. *Cantonese: A Comprehensive Grammar*. London: Routledge.
McMahon, April M. S. and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.
McMahon, April M. S., Paul Heggarty, Robert McMahon and Warren Maguire. 2007. The sound patterns of Englishes: Representing phonetic similarity. *English Language and Linguistics* 11: 113–142.
Mufwene, Salikoko S. and Stefan T. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30/1: 27–51.
Mukherjee, Joybrato. 2015. Responses to Davies and Fuchs. *English World-Wide* 36/1: 34–37.
Mukherjee, Joybrato and Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27/2: 147–173.
Mukherjee, Joybrato and Stephan T. Gries. 2009. Collostructional nativisation in New Englishes: Verb-construction associations in the International Corpus of English. *English World-Wide* 30/1: 27–51.
Nelson, Gerald. 2015. Responses to Davies and Fuchs. *English World-Wide* 36/1: 38–40.
Nordhoff, Sebastian. 2009. *A Grammar of Upcountry Sri Lanka Malay*. Utrecht: LOT.
Omar, Asmah H. and Rama Subbiah. 1989. *An Introduction to Malay Grammar*. Kuala Lumpur: Dewan Bahasa dan Pustaka.
Pandharipande, Rajeshwari V. 1997. *Marathi*. London: Routledge.
Peters, Pam and Tobias Bernaisch. 2022. The current state of research into linguistic epicentres. *World Englishes* 41: 320–332.

Platt, John, Heidi Weber and Ho Mian Lian. 1984. *The New Englishes*. London: Routledge and Kegan Paul.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.
Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7/2: 149–182.
Rohdenburg, Günter. 2003. Cognitive complexity and horror aequi as factors determining the use of interrogative clause linkers in English. In Günter Rohdenburg and Britta Mondorf eds. *Determinants of Grammatical Variation in English*. Berlin: Mounton de Gruyter, 205–250.
Romasanta, Raquel P. 2017. Contact-induced variation in clausal verb complementation: The case of REGRET in World Englishes. *Alicante Journal of English Studies* 30: 121–147.
Romasanta, Raquel P. 2019. Variability in verb complementation: Determinants of grammatical variation in indigenized L2 varieties of English. In Hanna Parviainen, Mark Kaunisto and Päivi Pahta eds. *Corpus Approaches into World Englishes and Language Contrasts*. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/20/romasanta/
Romasanta, Raquel P. 2021. Substrate language influence in Postcolonial Asian Englishes and the role of transfer in the complementation system. *English Studies* 102/8: 1151–1170.
Santoro, Gerald M. 1995. What is computer-mediated communication? In Mauri P. Collins and Zane L. Berge eds. *Computer Mediated Communication and the Online Classroom*. Cresskill: Hampton, 11–27.
Schachter, Paul and Fe T. Otanes. 1972. *Tagalog Reference Grammar*. Berkeley: University of California Press.
Schiffman, Harold F. 1999. *A Reference Grammar of Spoken Tamil*. Cambridge: Cambridge University Press.
Schneider, Edgar W. 2007. *Postcolonial English: Varieties around the World*. Cambridge: Cambridge University Press.
Schneider, Edgar W. 2012a. Contact-induced change in English worldwide. In Terttu Nevalainen and Elizabeth C. Traugott eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 572–581.
Schneider, Edgar W. 2012b. Exploring the interface between World Englishes and second language acquisition – and implications for English as lingua franca. *Journal of English as a Lingua Franca* 1/1: 57 – 91.
Schneider, Edgar W. 2013. English as a contact language: The 'New Englishes'. In Daniel Schreier and Marianne Hundt eds. *English as a Contact Language*. Cambridge: Cambridge University Press, 131–148.
Schreier, Daniel, Marianne Hundt and Edgar W. Schneider eds. 2019. *The Cambridge Handbook of World Englishes*. Cambridge: Cambridge University.
Setter, Jane, Cathy S. P. Wong and Brian H. S. Chan. 2010. *Hong Kong English*. Edinburgh: Edinburgh University Press.
Slobin, Dan. 1973. Cognitive prerequisites for the development of grammar. In Charles Ferguson and Dan Slobin eds. *Studies in Child Language Development*. New York: Holt, Rinehard and Winston, 175–208.
Slobin, Dan. 1977. Language change in childhood and history. In J. Macnamara ed. *Language Thought and Language Learning*. New York: Academic Press, 185–214.

Slobin, Dan. 1980. The repeated path between transparency and opacity in language. In Ursula Bellugi and M. Studdert-Kennedy eds. *Signed and Spoken Language: Biological Constraints on Linguistic Form*. Winheim: Verlag Chemie, 229–243.

Steger, Maria. 2012. *New Englishes are Simpler, isn't it? Morphosyntactic Iconicity in Institutionalized Second-Language Varieties of English*. Regensburg: University of Regensburg dissertation.

Steger, Maria and Edgar W. Schneider. 2012. Complexity as a function of iconicity: The case of complement clause constructions in New Englishes. In Bernd Kortmann and Benedikt Szmrecsanyi eds. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Berlin: Mouton de Gruyter, 156–191.

Strevens, Peter. 1980. *Teaching English as an International Language: From Practice to Principle*. Oxford: Pergamon Press.

Szmrecsanyi, Benedikt. 2013. Typological profile: L1 varieties. In Bernd Kortmann and Kerstin Lunkenheimer eds. *The Mouton World Atlas of Variation in English*. Berlin: Mouton de Gruyter, 826–843.

Szmrecsanyi, Benedikt and Bernd Kortmann. 2009a. Between simplification and complexification: Non-standard varieties of English around the world. In Geoffrey Sampson, David Gil and Peter Trudgill eds. *Language Complexity as an Evolving Variable*. Oxford: Oxford University Press, 64–79.

Szmrecsanyi, Benedikt and Bernd Kortmann. 2009b. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua* 119/11: 1643–1663.

Szmrecsanyi, Benedikt and Christoph Wolk. 2011. Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics* 11/2: 561–592.

Szmrecsanyi, Benedikt and Jason Grafmiller. 2023. *Comparative Variation Analysis: Grammatical Alternations in World Englishes*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt and Melanie Röthlisberger. 2019. World Englishes from the perspective of dialect typology. In Daniel Schreier, Marianne Hundt and Edgar W. Schneider eds., 534–558.

Tan, Ying-Ying. 2014. English as a ‘mother tongue’ in Singapore. *World Englishes* 33/3: 319–339.

Tegey, Habibullah and Barbara Robson. 1996. *A Reference Grammar of Pashto*. Washington: Department of Education.

Thomason, Sarah G. 2001. *Language Contact: An Introduction*. United States: Mouton de Gruyter.

Werner, Valentin. 2014. *The Present Perfect in World Englishes: Charting Unity and Diversity*. Bamberg: University of Bamberg Press.

Wheeler, Benjamin, Robert Englebretson and Carol Genetti eds. 2005. *Complementation in Colloquial Sinhala: Observations on the Binding Hierarchy*. Santa Barbara: University of California, Santa Barbara.

Williams, Jessica. 1987. Non-native varieties of English: A special case of language acquisition. *English World-Wide* 8/2: 161–199.

Yegorova, Raisa Petrovna. 1971. *The Sindhi Language*. Moscow: Nauka Publishing House.

Corresponding author
Raquel P. Romasanta
University of Santiago de Compostela
Facultade de Filoloxía
Avda. de Castelao, s/n
15782 Santiago de Compostela
Spain
E-mail: raquel.romasanta@usc.es

received: November 2023
accepted: February 2024

RiCL Research in Corpus Linguistics

Review of Timofeeva, Olga. 2022. *Sociolinguistic Variation in Old English: Records of Communities and People*. Amsterdam: John Benjamins. ISBN: 978-9-027-21134-7. DOI: https://doi.org/10.1075/ahs.13

James M. Stratton
Pennsylvania State University / USA

This book lays the foundations for studying the sociolinguistics of Old English, a period that has scarce sociolinguistic metadata and mass fragmentary evidence. While the Old English record is transmitted predominantly through a biased upper-class male religious elite, the author, Timofeeva, successfully illustrates that Old English has much to offer in the way of sociolinguistic information. Through eight chapters, the author surveys the applications of sociolinguistic methods and theories to Old English, focusing, in particular, on the application of social networks, and the extraction of sociodemographic information embedded in legal records.

Chapter 1 contextualizes the overall aim of the book by arguing that Old English can provide a glimpse into the social world of its speakers. While the dearth of sociolinguistic metadata may disable the possibility of using the composite extant record of Old English to reconstruct the social forces influencing Old English, the author maintains that if analysts circumscribe their analysis to a smaller subset of texts, Old English can still be sociolinguistically informative. The chapter begins with a discussion of major milestones in the reconstruction of extralinguistic influences on the history and development of Old English. The author subsequently turns to a review of two branches of sociolinguistics: correlational sociolinguistics and interactional sociolinguistics. Justifiably, the author raises issues concerning the representativeness of Old English texts and summarizes the genres and text categories available. This treasure trove of information will be of immediate interest to students and scholars of Old English, linguists, language historians, and Medievalists.

Chapter 2 illustrates how social network analysis can be applied to the study of historical documents and the texts associated with the court of King Alfred in the ninth century. King Alfred is revered as one of the most influential kings in British history, most notably, for his resistance

against the Vikings. Linguistically, Alfred is viewed as significant for his advocation for educational form, which led to an abundance of Old English records in the West Saxon dialect, abundant at least relative to the other attested regional varieties (Kentish, Mercian, Northumbrian). While many texts in Old English are anonymous and lack (scribal) authorship, Timofeeva argues that Alfredian texts can be used to reconstruct social networks, as these texts often contain scribal authorship information. The author subsequently illustrates how specific lexical choices (e.g., *Angelcynn*, *here*) appear to correlate with particular social networks (e.g., the court of Alfred), showing how social network analysis can be used to localize speech/scribal communities in historical periods.

Chapter 3 discusses the use of legal documents when studying or reconstructing the sociology of Old English. While the author states that letters "are commonly considered the best type of data" (p. 52) in historical sociolinguistics, she argues that legal documents can be used as a workaround for the absence of such texts in Old English. Since charters are a type of legal document that were designed to be read aloud, they may provide insight into oral language during this period. Analyzing more speech-related texts of this kind can be valuable to researchers interested in language change since it is often assumed that the locus of linguistic change is in spoken as opposed to written language (Milroy 1992: 32), with some honorable exceptions (e.g., Hinrichs and Szmrecsanyi 2007: 441; Jankowski 2013: 103–105). The author illustrates that, given their status as legal documents, charters contain names (e.g., recipients, witnesses), occupations (e.g., bishops, clerics), titles (e.g., *eorl*, king), and information about gender —crucial extralinguistic information which can be used to reconstruct the sociohistorical sphere of Old English. At the end of the chapter, the author references four major charter types, which sets up the structure of the following four chapters: Chapter 4 (diplomas), Chapter 5 (writs), and Chapter 6 (wills).

Of the chapters that follow, in my view, Chapter 5 is most insightful, as it illustrates how linguistic variation (i.e., salutation choice between *freondlice* vs. *eadmodlice*) appears to correlate with the "social status" of the sender and addressee. This finding suggests, in line with the Uniformitarian Principle (Labov 1972: 275), that the forces at play today (e.g., socioeconomic status and power) were also likely at play historically. The analysis of wills in Chapter 6 is also particularly valuable for its insights into differences between male and female language. In an analysis of cursing, the author reports that women who had wills prepared for them used more cursing than men who had wills prepared for them, speculating that the higher use of cursing was due to women's predilection for being more "emotional". Since there is a long history of discourse and public perception that women are more emotional in their language (e.g., Stoffel 1901: 101–102; Peters 1994), an observation that still appears to hold true for some linguistic variables today

(e.g., Tagliamonte and Roberts 2005; Tagliamonte 2008), it is fascinating to see some potential evidence that this was also true historically.

Chapter 7 examines changes in the community of practice of the royal chancery and in the discourse community of the local courts in Early Middle English. Examining loanwords in Latin legal documents, the author shows how the borrowings illustrate a change in the scribal authorship, moving from Anglo-Saxon scribes to Norman scribes. Norman linguistic influence was strongest at the verbal domain but was also prevalent in inflectional morphology. In a short epilogue, in Chapter 8 the author returns to the principal aim of the book which was to explore "the possibilities of a sociolinguistic enquiry into the Old English period" (p. 175). The author certainly achieved this objective, showcasing new methodological and qualitative procedures.

If I may indulge in any criticism, with advances in recent decades in statistical methods in sociolinguistics and the notable shift in historical linguistics toward a more quantitative science (Jenset and McGillivray 2017; Brinton *et al*. 2021; Kortmann 2021), this book could have benefited from the application of advanced statistical methods, in particular mixed effects modeling, to uncover "orderly heterogeneity" (Weinreich *et al*. 1968). Mixed effects regression models have become a cornerstone of correlational sociolinguistics and studies have illustrated the application of these models to Old English data (De Cuypere 2015; Stratton 2022, 2023). However, the underlying goal of the book was clearly to bring forth new methodologies and approaches and illustrate how modern sociolinguistic theories may be applied and tested in Old English. It goes without saying that the author certainly achieved this goal, and readers will gain a great deal from the case studies and application of sociolinguistic theories and methods addressed in this book.

This work provides an important first step towards unpacking the complex sociolinguistic makeup of Old English. While the number of extant Old English texts will likely remain stable in the coming years, analysts can invigorate the data with new methods and theories and can turn to this body of work for inspiration. This book will inevitably be of great use to students and scholars of Old English, historical (socio)linguists, and language historians, and possibly even legal scholars.

REFERENCES

Brinton, Laurel, Patrick Honeybone, Bernd Kortmann and Elena Seoane. 2021. 25 years of English Language and Linguistics: A celebration and analysis. *English Language & Linguistics* 25/4: 677–685.

De Cuypere, Ludovic. 2015. A multivariate analysis of the Old English ACC+ DAT double object alternation. *Corpus Linguistics and Linguistic Theory* 11/2: 225–254.

Hinrichs, Lars and Benedikt Szmrecsanyi. 2007. Recent changes in the function and frequency of standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics* 11/3: 437–474.
Jankowski, Bridget L. 2013. *A Variationist Approach to Cross-register Language Variation and Change*. Toronto, University of Toronto dissertation.
Jenset, Gard B. and Barbara McGillivray. 2017. *Quantitative Historical Linguistics. A Corpus Framework*. Oxford: Oxford University Press.
Kortmann, Bernd. 2021. Reflecting on the quantitative turn in linguistics. *Linguistics* 59/5: 1207–1226.
Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.
Milroy, James. 1992. *Language Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.
Peters, Hans. 1994. Degree adverbs in Early Modern English. In Dieter Kastovsky ed. *Studies in Early Modern English*. Berlin: Mouton de Gruyter, 269–288.
Stoffel, Cornelis. 1901. *Intensives and down-toners. A study in English adverbs.* (Anglistische Forschungen. Heft 1). Heidelberg: Carl Winter's Universitätsbuchhandlung.
Stratton, James. 2022. Old English intensifiers. The beginnings of the English intensifier System. *Journal of Historical Linguistics* 12/1: 31–69.
Stratton, James. 2023. Where did *wer* go? Lexical variation and change in third-person male adult noun referents in Old and Middle English. *Language Variation and Change* 35/2: 199–221.
Tagliamonte, Sali A. and Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series Friends. *American Speech* 80/3: 280–300.
Tagliamonte, Sali A. 2008. So different and pretty cool! Recycling intensifiers in Canadian English. *English Language and Linguistics* 12/2: 361–394.
Weinreich, Uriel, William Labov and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. Directions for historical linguistics. In Winfred P. Lehmann and Yakov Malkei eds. *Directions for Historical Linguistics*. Texas: University of Texas Press, 95–195.

Reviewed by
James M. Stratton
Pennsylvania State University
Department of Germanic and Slavic Languages and Literatures
442 Burrowes Building
University Park, PA
16802
e-mail: james.stratton@psu.edu

RiCL Research in Corpus Linguistics

Review of Crosthwaite, Peter ed. 2024. *Corpora for Language Learning: Bridging the Research-Practice Divide*. London: Routledge. ISBN: 978-1-032-53722-1. DOI: https://doi.org/10.4324/9781003413301

Mohammad Ahmadi
Lorestan University / Iran

Corpus Linguistics (CL) is an important field of applied linguistics that has enriched the investigation of language in use to a great extent. Today, CL has found application in several areas and has paved the way to new vistas in writing, second language acquisition, lexicography, and related fields. Thus, corpus application is most apparent in the enhancement of writing production since textual analysis of a learner's texts can enlighten such matters as mistakes, tendencies for preferred collocations, or any other questions whenever relevant to a certain topic. Evidently, CL owes significant debt to the work of leading researchers such as John Sinclair (1991) and Susan Hunston (2002), who have revolutionized the quality of language teaching and learning. They have particularly highlighted the fact that data obtained from CL needs to be integrated into language teaching/learning. Their contributions also led to the emergence of specific corpora, carefully tailored to address particular needs of educators, thereby enhancing language teaching and learning (see Römer 2010).

Nevertheless, Peter Crosthwait's edition *Corpora for Language Learning: Bridging the Research-Practice Divide* has stressed the challenges of incorporating CL into the learning of language. Probably the biggest one is the lack of connection between the latest corpus research and its real-life use in the classroom. This resource is crucial as many educators may be unfamiliar with corpus tools or lack adequate training in data-driven learning (DDL). The volume also elaborates on how the field of corpus research should be integrated into the practices of language teaching. These issues are addressed by providing theoretical concepts and practical guidelines for DDL implementations across

different fields and levels of education, with contributions from renowned international scholars.

The book is structured into 17 chapters, allowing readers to navigate through different sections and follow a coherent sequence of ideas. Initially, Crosthwaite outlines the organization of the book by noting that each chapter is followed by some discussion sections reflecting the perspectives of researchers, teachers or learners who were influenced by that work. The discussions are not in the form of typical research articles; rather, they take the form of personal reflections. In the first chapter, Crosthwaite explains the significance of bridging the research-practice divide in DDL, stressing the need for better connections between academic theories and teaching methods. This is then succeeded by an interview with Laurence Anthony in Chapter 2 where he describes the 2021 update of *AntConc*, which has been redesigned in Python with an SQLite database for proper optimization and improvement. Commenting on the practicality and user-friendliness of *AntConc*, Anthony shows how learners use *AntConc* to do DDL through finding the lexical units by word/keyword lists and n-gram, modifying and distributing new corpora, handling large volumes of data and using effective statistical evaluation. As he notes, "the greatest challenge in DDL for learners is simply finding the target texts and loading them into the concordance" (p. 13).

Chapter 3 is devoted to the use of multimodal corpus data and analytical approaches in language education to improve students' multisemiotic approaches to meaning-making. Tony Berber Sardinha explains the various ways in which computers can systematically describe images using computer vision techniques, including *Google Cloud Vision Application Programming Interface*. One critical point which is covered in this chapter is the practice of using multidimensional analysis through a particular statistical procedure, known as canonical correlation, which is a corpus linguistics technique, "to detect the dimensions from one particular mode that align with the dimensions from another mode" (p. 28). This allows exploring "discourses, ideologies, and visual content that shape social media conversations" (p. 34).

Chapter 4 is a conversation with Alex Boulton, a DDL expert. According to the text, DDL entails getting students to learn language patterns from corpus data without actually being taught. "They do this not by learning 'rules' but by looking at how language is actually used" (p. 43). In the discussion, Boulton specifies the history of DDL, advantages, disadvantages, recent innovations, theoretical implications, technological

advances, and its application to skills beyond writing. He also recommends that greater efforts should be put into the assessment of DDL usability and encourages teachers to introduce DDL activities in the classroom.

The issue of employing DDL methodologies in languages other than English (LOTEs), particularly the L2 context, is the main focus of Chapter 5 authored by Luciana Forti. It can be noted, however, that DDL has been successfully applied for English only, although the author enumerates some possibilities to enhance interaction between DDL, LOTE practices, and SLA theories. The chapter highlights a DDL study on polysemous Italian words and encourages further research to expand DDL application to other LOTE contexts.

In Chapter 6, Ana Frankenberg-Garcia examines a way of applying DDL and corpus tools to enhance learners' appreciation of Academic English. She presents *ColloCaid*, a web-based DDL tool created with academic collocations which enables users to search for collocations, disambiguate words, view concordances and a collocational network, and see example sentences. It is worth noting that the users should not expect the features offered by a professional text editor from *ColloCaid*; rather, it is mainly a "proof-of-concept tool that provides academic English collocation suggestions" (p. 74).

Chapter 7 is devoted to the progression of incidental acquisition through a framework which involves extensive reading (ER) and extensive viewing (EV). Referring to CL findings, Clarence Green stresses the role of comprehensible input in the development of vocabulary and collocations in particular. He further emphasizes the importance of multimedia annotation technology to enhance the comprehensibility of input. Looking at the text in terms of vocabulary difficulty by employing corpus tools, Green recommends that appropriate extensive reading and viewing material should be chosen.

In Chapter 8, Reka R. Jablonkai provides an overview of three main approaches to corpus-based pedagogy, namely, corpus-informed teaching, integrated corpus-supported teaching and learning, and self-directed DDL. Then, she discusses general information about DDL, theoretical background, and a pedagogical model. The chapter offers an insight into various DDL activities, teaching of collocations, lexical phrases as well as discipline specific lexical items.

Chapter 9 sees Tatyana Karpenko-Seccombe looking into the means of adding corpus tools and DDL to enhance the students' argumentation in the academic writing process. The author presents teaching recommendations on corpus consultations, argumentation, patterns of claims and supports, and problem-solution patterns. DDL activities highlighted in the chapter are concerned with such tasks as the concordancers' use to compare collocations, to analyze the distribution of the term across disciplines, or to undertake research with corpora. Focus is given to the tools like *SkELL*, *Lextutor*, and *MICUSP*.

Chapter 10, which is an interview with Tove Larsson and Douglas Biber, cautions against exclusive use of statistical indicators and opaque calculations in CL. It stresses the need for linguistic interpretability and accuracy in research methodology. The chapter further illustrates the peril of relying on quantitative counts without access to annotated texts which can be "problematic if we cannot assess the accuracy of the output" (p.135). The authors also call for a linguistically-motivated paradigm for the analysis of corpora.

Elen Le Foll, in Chapter 11, participates in an interview where she supports open science and education on the interface of CL with language teaching. She especially pays great attention to the issues of openness to knowledge, information sharing, and cooperation. The recommendations include providing free access to research papers, corpus data, and tools, and emphasizes the need for addressing the separation of research and practice in teaching.

In Chapter 12, Agnieszka Leńko-Szymańska has considered how corpora and CL could be applied to assessing learners' L2 vocabulary knowledge. It discusses the issues involved in measuring the extent of word knowledge stressing on the fact that it is not easily measurable since it has many dimensions. The chapter also shows that corpora offer actual language data, and it is possible to develop vocabulary tests on their base. The chapter also discusses the advantages of getting direct access to corpus data and the application of learner corpus data in assessment and modelling of vocabulary.

Chapter 13 is primarily devoted to discussing CL as one of the components of teacher education programs. Qing Ma offers a two-step training framework for the implementation of corpora into teacher education programs for both pre-service and in-service teachers. The challenges and strategies for implementation are described in the chapter, and empirical studies concerning the outcomes of the corpus-based instruction

are outlined. These are followed by the subsequent appeal for more research as to the effects of the range of corpus-based procedures on teacher knowledge and practice.

Chapter 14 focuses on how DDL, CL technology, and phraseography can help improve learners' knowledge of collocation and multiword units. More specifically, Adriane Orenha-Ottaiano presents DDL, describes the activities of DDL on corpora and concordancing tools and focuses on the demand for the accurate frequency of data from corpora. The author also expands the specifics of phraseology and offers corpus-based development activities to be included in the materials.

Chapter 15 ventures into examining broad data-driven learning (BDDL) and its potential when applied to the process of learning informal language supported by technology. Pascual Pérez-Paredes criticizes existing DDL approaches and proposes augmenting DDL to support self-initiated, self-managed learning, utilizing learners' personal electronic data. The author further illustrates examples of tools and resources for BDDL and recommends employing Natural Language Processing and machine learning techniques, as well as, the DDL tasks for the informal settings.

Chapter 16 is concerned with using DDL and CL in the context of EAP when enhancing the internationalization of higher education. Paula Tavares Pinto focuses largely on how DDL and corpus-based activities complement each other, emphasizing their application in analyzing subject-specific academic corpora, teaching academic language patterns, and raising awareness of variations in academic register. The chapter also puts forward the suggestion for the use of the corpora and corpus tools, advantages and limitations of the application of DDL in EAP settings, and the necessity of integrating DDL approaches into the practices in order to assist multilingual scholars and to provide them with the tools to engage in English-medium academic debates.

Chapter 17 of the current volume is specifically devoted to the convergence between CL and EAP. Vander Viana elaborates on the benefits that corpora as well as the corpus-based approaches bring about in EAP research and application. Viana examines how these methods can be used to investigate academic language rigorously, moving beyond intuition or idealized rules. The chapter illustrates methods including compiling specialized academic corpora for the study, genre analysis of texts, cross-comparison of expert and learner writing using learner corpus analysis, and analysis of variations in register across the disciplines using multi-dimensional analysis. Viana further highlights

the significance of enthusiastically produced and selected EAP corpora and also focuses on the interconnection between corpusers (CL researchers) and EAP specialists.

To conclude, the book contains a lot of useful information for the practical implementation of corpus-assisted language learning. However, it could benefit from more critical evaluation and workable solutions for tackling challenges commonly associated with the integration of DDL, and DDL tools into mainstream education. Although the current resource recognizes the problems and potential drawbacks of DDL, it might not sufficiently delve into these issues or provide solutions for addressing them. Moreover, it might not adequately address such issues as pedagogical concerns regarding the relevance of corpus data to specific learning contexts and the potential for these tools to distract from other important aspects of language learning (see Boulton and Cobb 2017).

The book also introduces corpora in vocabulary assessment but fails to discuss other dimensions of the assessment exhaustively. For example, the current coverage could have been improved by articulating how corpora are used in making decisions about grammatical error, pragmatic competence, or specific discoursal features. More exemplifications of assessment tasks which are based on the corpus approach, for instance, scrutinizing the texts written by learners with regard to particular linguistic characteristics or devising performance assessments grounded on the actual language data, would contribute to the elaboration of the issue. The expansion of the assessment perspective would offer a better perception of corpora's contribution to the evaluation of language learning results.

Apart from the issues pointed out in the previous section on assessment coverage, this volume also lacks detailed guidance on teacher training in corpus-based language pedagogy (CBLP). While the book advocates the importance of teacher training in CBLP and introduces a two-step framework for developing corpus literacy and pedagogical skills, it did not provide detailed instruction on how to implement it appropriately. In the absence of proper training, the teachers are likely to struggle in implementing some of these novel measures in their teaching programs, leading to suboptimal outcomes for students. For example, in various parts of the book, there are references to the necessity of teacher education. A statement like

> CL-literate EAP practitioners are possibly more capable of designing a better endowed exploration of content for discipline-specific EAP, which has implications for teacher education and professional development practices as well (p. 255)

highlights the importance of teacher training but the related sections do not specify a detailed strategy for accomplishing it.

Nonetheless, *Corpora for Language Learning* serves as a useful starting point for those with scholarly interest in CL and DDL. Teachers and scholars in the field of language learning and teaching can learn more about the theoretical framework of the corpus-based language instruction and can get a much deeper insight into the relevant DDL instruments such as *AntConc*, *WordSmith* or *CorpusMate*, as well as the ways for their appropriate use in language learning and teaching. Overall, this resource is very valuable for anyone intending to improve their knowledge in the area of data-driven language education in different settings.

References

Boulton, Alex and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67/2: 348–393.

Hunston, Susan. 2022. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Römer, Ute. 2010. Using general and specialized corpora in English language teaching: Past, present and future. In Mari Carmen Campoy, Begoña Bellés-Fortuno and María Lluïsa Gea-Valor eds. *Corpus-based Approaches to English Language Teaching*. London: Continuum, 18–35.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Reviewed by
Mohammad Ahmadi
Department of English Language
Lorestan University
Khorramabad, Lorestan Province
68151-44316
Iran
e-mail: ahmadi.m@lu.ac.ir

RiCL Research in Corpus Linguistics

Review of Barth, Danielle and Stefan Schnell. 2022. *Understanding Corpus Linguistics*. London: Routledge. ISBN: 978-0-429-26903-5. DOI: https://doi.org/10.4324/978042926903

Isabel Zimmer – Elen Le Foll
University of Cologne / Germany

Understanding Corpus Linguistics provides an introduction to corpus linguistics and its application in multiple areas of linguistics. Written with undergraduate and graduate students of linguistics in mind, the textbook outlines how corpora can improve our knowledge of languages by providing authentic data. The book is divided into three parts: the first part provides an overview of basic concepts of corpus linguistics, the second part focuses on the processes of working with and creating corpora, while the third part explores the contribution of corpora to a selection of sub-disciplines of linguistics.

The opening chapter outlines the focus of corpus linguistics by differentiating it from other methods used in linguistics, such as acceptability judgments and experimental setups. It weighs up the advantages and disadvantages of corpus-linguistic methods. Barth and Schnell further illustrate the diverse applications of corpus linguistics in various sub-disciplines of linguistics, including sociolinguistics, psycho- and neurolinguistics.

Chapter 2 provides a comprehensive overview of the basic terminology of corpus linguistics, for instance, elaborating on the difference between word forms and lexemes, and types and tokens. It briefly explains how situational and/or text-internal contexts can influence the use of linguistic forms. Furthermore, the authors address the fact that corpora can only ever reflect a subset of language use, sampled from a relatively limited portion of the population.

The third chapter focuses on corpus composition and corpus types. Barth and Schnell emphasise that representativeness and a balanced number of text types are essential for corpora as an empirical data basis, and further highlight the importance of metadata, hereby highlighting potential issues with web corpora.

The objective of Chapter 4 is to demonstrate that corpus linguistics can be applied to many subdisciplines, for instance for questions in morphology or phonetics, discourse or sign language. For each subdiscipline, one or two studies are presented.

Although Chapter 5 is entitled “Corpus Queries”, it does not introduce queries as such, but rather different methods for analysing corpus data. This is likely explained by the fact that the authors refrain from explaining the use of any specific tool or programming package to avoid the textbook quickly becoming outdated. That said, the chapter does list some example corpus tools, *R* packages, and *Python* modules in standalone textboxes, which help to concretise the explanations. There is a strong focus on quantitative corpus linguistic methods with surprisingly little emphasis on concordancing, which is mentioned only briefly after frequency lists, keywords, dispersion plots, Zipfian distributions, collocations and bigrams, and association measures. The chapter ends with an introduction to regular expressions for corpus querying.

Chapter 6 begins by introducing three corpus building scenarios that allow the authors to cover a range of typical situations in the space of a short chapter. The chapter addresses the identification, selection, and evaluation of texts for corpus compilation, the collection procedure, copyright and privacy, as well as technical considerations such as how to deal with different modes and scripts. It provides more detailed information about the transcription of spoken data, including how to link transcripts to raw data using *ELAN* (2020) and concludes with short sections on data formats, the inclusion of metadata, and how to publish a corpus.

In Chapter 7, which deals with corpus annotation, Barth and Schnell provide brief descriptions of different types of annotation, from phonetic and prosodic annotation to discourse and reference annotation, including morphological and semantic annotation and part-of-speech (POS) tagging. The second half of the chapter focuses on corpus annotation in the context of cross-linguistic typological research with concrete examples from *The Social Cognition Parallax Interview Corpus* (SCOPIC; Barth and Evans

2017) and the *Multilingual Corpus of Annotated Spoken Texts* (Multi-CAST; Haig and Schnell 2015).

Chapter 8 begins by introducing foundational concepts of statistics for corpus linguistics including sampling, dependent and independent variables, distributions, range, and spread. It continues with worked examples of the use of chi-square and correlation tests, mixed-effects logistic regression, classification tree, and random forests applied to real-life corpus data. It also includes shorter sections on clustering and on how to report the results of statistical tests.

Following a brief introduction to sociolinguistics, Chapter 9 explains how corpora are used in the study of dialect and regional variation and dialectometry. It outlines the types of variables typically included in sociolinguistic studies and how corpus analyses can inform our understanding of variation and language change.

Chapter 10 focuses on language documentation and its use of corpora. Despite the smaller size and limitations of many corpora in this field, the authors emphasise their crucial role in preventing language loss. They elaborate on the process of corpus building in language documentation and provide examples of annotations of different data types. Due to the limited size of the corpus, research questions must be adapted, and only specific objectives can be addressed, in contrast to those that can be analysed with larger corpora. Despite the limited corpus size, the authors exemplify different analyses that can be conducted and conclude this chapter by discussing the limitations and advantages of smaller corpora.

Chapter 11 introduces corpus-based typology. The authors discuss some of the assumed language universals, demonstrating that there is, in fact, considerable variation across different languages. Linguistic diversity is exemplified with the use of different expressions for referential entities. To conclude, the authors outline issues and biases in corpus-based typology research, particularly when working with corpora that consist mainly of written data.

Understanding Corpus Linguistics is written in accessible language. Even relatively basic terminology is explained in detail so that beginners —the target readership of the textbook— are well catered for in this respect. The authors cover a lot of material in a relatively short textbook. Each chapter begins with a list of keywords

and concludes with some recommended further readings. The exercises, which build on what has been previously explained, are also a great addition to the textbook.

Whilst we recognise that there is no perfect order that will suit all readers, we found that the order of some of the chapters was not the most intuitive for us. In particular, we believe that the order of Chapters 5–7 (corpus queries → corpus building → corpus annotation) may be difficult for corpus novices. Whichever order is chosen (and the chosen order certainly has its justifications), cross-references between the chapters would help the reader to find the information they need more easily. A real strength of the textbook is that it mentions many different corpora, representing a wide range of languages and designed for use in different subdisciplines of linguistics.

As the title suggests, this textbook is about *understanding* corpus linguistics, not necessarily about *doing* corpus linguistics. As a result, Chapter 7, for example, focuses explicitly on the types of annotation and annotation schemes used in corpus linguistics, independently of any software or tool. With this in mind, this focus on tagsets rather than POS taggers makes sense, but it does mean that the reader is left with no idea as to how to actually perform a task as simple as POS-tagging, which may prove frustrating for some.

Chapter 8, on statistics for corpus linguistics, contains some inaccuracies. On p. 138, it states that "parametric tests need to meet assumptions like following a normal distribution and independence. Otherwise, non-parametric tests need to be used." This statement suggests that non-parametric tests can be used when the assumption of the independence of the data points is violated. This is not the case for most non-parametric tests used in corpus linguistics, for which the independence assumption still holds (for instance, the Mann–Whitney U test or the Kruskal-Wallis test). Most problematically, on p. 152, we read that "the *p*-value represents the chance that the null hypothesis would be true if we observed this sample of data." This definition of *p*-values is incorrect. Although they are often misconstrued as such, *p*-values do not correspond to the probability that the null hypothesis is true or false (see, for instance, Winter 2020: 171). Rather, *p*-values correspond to the probability of observing a result as extreme as, or more extreme than, the one obtained from the sample, if the null hypothesis were true. The "Further Reading" section of this chapter lists several books that provide excellent introductions to statistics for linguistics (including Winter 2020) and that are at an appropriate level for the target readership of the textbook. In contrast, we fear that the

advice to "just start reading [online] forums and seeing what you can glean [...]" (p. 163) is less sensible and unlikely to lead to sound statistical literacy among student readers.

Throughout the textbook, the authors place a commendable emphasis on the reproducibility and replicability of corpus linguistics research. For instance, in Chapter 5, they stress the importance of documenting workflows, which is rarely mentioned in corpus linguistics textbooks. Barth and Schnell also innovate by mentioning two studies which some would consider to be 'null results' (Chapter 9), but which we agree are very much still worth reporting about. It is also refreshing to see that the authors include the publication of a corpus as a:

> definitional feature of any corpus [...] because the primary purpose of a corpus is serving as a source for linguistic research, and being data (literally meaning 'given') entails that language scientists should be given the opportunity to look at the same things (the 'data'), including the surrounding context, when evaluating linguistic analyses and respective theories (p. 93).

While this is a very welcome addition to the definition of a research corpus, the link between the section on the "Publication of the Corpus" (6.6) and on the "Availability of Texts: Copyright and Privacy" (6.6.2) could be made clearer. It feels somewhat of an understatement to claim, on page 97, that "even some texts available through the internet may have some copyright protection or restrictions on usage." Given that the idea of publishing corpora has not yet been fully embraced by the corpus linguistics community, it might be worth including some examples of repositories where corpora can be made available. The authors choose the example of the Multi-CAST corpus, which is worth mentioning for several reasons, but which is published on a dedicated corpus website. We have our doubts as to whether this is the best example of how to share a corpus. Corpus websites need to be maintained and, as many older projects have sadly shown, links quickly become broken, resulting in the loss of valuable corpus resources. In addition, building an entire website may seem overwhelming to many researchers. For both these reasons, we suggest mentioning open repositories —for instance, CLARIN,[1] OSF,[2] TrolLing[3] or Zenodo[4]— which provide more sustainable infrastructures for corpus sharing with quick and easy uploading procedures.

[1] https://www.clarin.eu/content/data
[2] https://osf.io/
[3] https://dataverse.no/dataverse/trolling
[4] https://zenodo.org/

In conclusion, *Understanding Corpus Linguistics* covers a lot of ground, while making complex concepts of corpus linguistics genuinely digestible for undergraduate and graduate students. Although we picked up on a few problematic passages, particularly in the statistics chapter, they do not overshadow the book's overall value. What sets this textbook apart is its commendable emphasis on providing examples in languages other than English, including signed languages, a very welcome addition to the existing corpus linguistics literature, which has so far had a very strong focus on English as the object of study. The textbook is also pioneering in its strong focus on typology, while at the same time offering interesting insights into how many other subdisciplines of linguistics can benefit from corpus research. By addressing these neglected areas, this textbook effectively fills a conspicuous gap in existing corpus linguistics textbooks, making it a valuable resource for linguistics students and educators alike.

References

Barth, Danielle and Nicholas Evans. 2017. SCOPIC design and overview. In Danielle Barth and Nicholas Evans eds. *LD&C Special Publication No. 12: The Social Cognition Parallax Interview Corpus (SCOPIC): A Cross-linguistic Resource*. Honolulu: University of Hawai'i Press, 1–23.

ELAN. 2020. (Version 6.0) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan.

Haig, Geoffrey and Stefan Schnell eds. 2015. *MultiCAST (Multilingual Corpus of Annotated Spoken Texts)*. https://multicast.aspra.uni-bamberg.de. (23 December 2024.)

Winter, Bodo. 2020. *Statistics for Linguists: An Introduction Using R*. London: Routledge.

Reviewed by

Isabel Zimmer
University of Cologne
CRC 1252 "Prominence in Language"
Luxemburger Str. 299
50939 Cologne
Germany
E-mail: ifuhrma2@uni-koeln.de

Elen Le Foll
University of Cologne
Department of Romance Studies
Universitätsstraße 22
50937 Cologne
Germany
E-mail: elefoll@uni-koeln.de

RiCL Research in Corpus Linguistics

Review of Loureiro-Porto, Lucía. 2024. *Pragmatic Markers in World Englishes:* Kind of *and* sort of *as a Case in Point*. València: Publicacions de la Universitat de València. ISBN: 978-8-411-18306-2. DOI: https://dx.doi.org/10.7203/PUV-OA-307-9

Sven Leuckert
Technische Universität Dresden / Germany

Pragmatics is one of the most vibrant fields in present-day linguistics, with an abundance of publications covering a broad range of phenomena and their representation in different linguacultural contexts. It comes as a bit of a surprise, then, that the pragmatic study of World Englishes has only become a major concern of the field somewhat recently (compared to, for instance, research on the morphosyntactic and phonetic properties of varieties of English). The monograph *Pragmatic Markers in World Englishes:* Kind of *and* sort of *as a Case in Point* by Lucía Loureiro-Porto is a new study in the growing body of publications using variationist tools to study pragmatic markers in different varieties of English. The author presents an empirical, corpus-based study of *kind of* and *sort of* across two 'Inner-Circle' varieties, British (BrE) and American English (AmE), and two 'Outer-Circle' varieties, Singapore English (SingE) and Philippine English (PhilE).

Chapter 1 offers an introduction both to the book's central framework as well as its aims and structure. On the very first page, the author points out that one of the central questions in World Englishes has been whether they are "becoming more similar or more distinct from each other" (p. 9). The focus is thus clearly on the notions of 'convergence' and 'divergence', two concepts of biggest importance in sociolinguistics. In addition to setting the thematic scope, the introduction also provides a brief overview of pragmatic markers, including both traditional and newer approaches, and a first insight into current studies into pragmatic markers in World Englishes. Finally, the introduction also outlines the monograph's aims and structure. The volume's main aim is "to provide a syntactic

and semantic-pragmatic characterization of the pragmatic markers *kind of* and *sort of*" (p. 16) in BrE, AmE, SingE, and PhilE.

Chapter 2 further sets the stage by discussing the origins and development of *kind of* and *sort of*. After an overview of the historical origins of the forms (Germanic for *kind* and Romance for *sort*), the main focus of the chapter is on the various linguistic processes that have affected the forms over time. Importantly, as emphasised by Hopper and Traugott (2003) and as established in the chapter, several important processes, such as reanalysis and grammaticalisation, are interconnected and cannot be teased apart neatly. Based on several aspects in the development of the two forms, however, "the diachronic evolution of *kind of* / *sort of* is considered as an illustration of grammaticalization" (p. 40).

In Chapter 3, the focus shifts to expressions of *kind of* and *sort of* in Inner-Circle contexts. After an overview of potential realisations and the semantic functions of *kind of* and *sort of*, a section is devoted to the pragmatic description of the two forms as stance markers and hedging devices. The chapter also explains the terminological choice of referring to the two constructions as 'pragmatic markers' (as opposed to one of the many other terminological options in use to describe these two and similar forms).

The focus then shifts to Outer-Circle contexts in Chapter 4. First, the chapter provides an overview of key models of World Englishes but explains that Kachru's (1985) Circles model is chosen as the dominant model due to its handy terminology and categorisation. This overview is followed by a survey of the four varieties in focus, BrE, AmE, SingE, and PhilE, and pragmatic differences between Inner- and Outer-Circle contexts. The term used to describe the study of pragmatics in World Englishes from a variationist angle is 'postcolonial pragmatics', which takes into account the fact that essential pragmatic concepts, such as face, may differ greatly between a given postcolonial society and a 'traditional', Inner-Circle context. The chapter closes with a brief overview of previous studies on pragmatic markers in SingE and PhilE.

Chapter 5 represents the core empirical chapter of the monograph. The methodological section at the beginning of the chapter introduces the data source for the analysis, the *Corpus of Global Web-based English* (GloWbE; Davies 2013). The author describes the corpus and its advantages but, importantly, also points out its drawbacks (such as the lack of insight into details about how the texts in the corpus were produced, and by whom). The section on the corpus is followed by a description of methodological

choices made for the analysis, such as how an appropriate sample was selected, and which variables are of importance in the analysis. This part leads into the actual analysis, starting with an overview of the overall results and frequency statistics before a closer look at the syntactic positions of *kind of* and *sort of* and a semantic and pragmatic analysis. A key finding of the pragmatic analysis confirms the "coexistence of the hedging and stance marker function" (p. 111), which appears to be prevalent in the dataset.

In Chapter 6, the processes of Americanisation and colloquialisation are employed to explain, at least to some extent, the tendencies identified in the data. Americanisation is taken as a promising candidate to interpret the closeness of the SingE to the AmE data, with the idea that SingE has, in a sense, 'caught up' compared to PhilE that historically originated as an AmE-based variety. Colloquialisation, in turn, affects language as "the process that refers to the tendency of the written language to incorporate features that are associated with the spoken conversational language" (p. 119; see Mair 1997). While both Americanisation and colloquialisation are attractive in explaining the findings, the chapter concludes with a final section that clarifies, importantly, that things are not that simple. In particular, the contrasting forces of global and local preferences need to be taken into account for a comprehensive picture.

Finally, Chapter 7 offers a conclusion and an outlook. After summaries of the volume's chapters, the author makes a plea for more diversified and larger datasets, in particular ones that include more varieties and, potentially, more samples for analysis.

The present monograph is a valuable contribution to the empirical study of pragmatic markers and offers both a detailed review of previous work and new findings derived from (admittedly complicated) data. Something to appreciate in particular is the care with which arguments are made in the book. While the author is clear about the direction of her research and the interpretation of the findings, she is also careful about leaving room for alternative ways of looking at the investigated constructions and why their frequencies and properties may differ across varieties. This carefulness is also applied to GloWbE as a corpus, which contains notoriously tricky data (see, for instance, the summary in Shakir and Deuber 2023). However, I would also argue that —while there are many things to praise about the title— there are also some aspects that hold it back in certain ways. My main criticism concerns the book's composition and overall structure: a substantial amount of space is devoted to fleshing out the backdrop of the study, including the historical origins of *kind of* and *sort of* and their functions, but also models

of World Englishes and so forth. That is not a problem in itself, but at ca. 130 pages of running text, the book is comparatively short and more akin to recent shorter publication formats, such as Cambridge Elements, than it is to more typical full-length monographs. My impression is that the book sits somewhat uncomfortably between a lengthier article and a fully-fledged monograph, which could have perhaps been dealt with by reorganising some of the chapters. This impression is exacerbated by the great level of detail given to background literature in the earlier chapters of the book but the at times slightly less detailed methodology and results section. For instance, on page 84, the author introduces the variables that are supposed to be analysed later on, but we learn about them in more detail only as part of the analysis and not in the methodological section itself. I was also wondering about the internal structure of some chapters, in particular Chapter 3, which, per its title, is about *kind of* and *sort of* in Inner-Circle Englishes, but doubles in function as an introduction to the semantic and pragmatic dimensions of the two markers (which seems odd, since those dimensions are clearly also relevant to other varieties). Apart from these more structural concerns, there is also a noticeable number of editing issues and typos (such as *many* instead of *may* on p. 11 or a full stop that became part of a footnote on p. 69), but these do not take away from the study in any meaningful way.

Overall, despite some caveats, I would certainly recommend this study to anyone with an interest in pragmatic variation in World Englishes. The substantial level of detail provided on *kind of* and *sort of* is extremely useful as an overview and the empirical analysis is solid and offers another small piece to the puzzle of pragmatic variation across varieties of English.

References

Davies, Mark. 2013. *Corpus of Global Web-based English*. https://www.english-corpora.org/glowbe/ (27 April, 2025.)

Hopper, Paul J. and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge: Cambridge University Press.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk and H. G. Widdowson eds. *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Mair, Christian. 1997. Parallel corpora: A real-time approach to the study of language change in progress. In Magnus Ljung ed. *Corpus-based Studies in English*. Amsterdam: Rodopi, 195–209.

Shakir, Muhammad and Dagmar Deuber. 2023. Compiling a corpus of South Asian online Englishes: A report, some reflections and a pilot study. *ICAME Journal* 47/1: 119–139.

Reviewed by

Sven Leuckert
Technische Universität Dresden
Institute of English and American Studies
Wiener Str. 48, room 3.10
01219 Dresden
Germany
E-mail: sven.leuckert@tu-dresden.de