far do web registers represent the linguistic characteristics of spoken registers? They note that the relationship between colloquial written registers and speech has long been of interest (it was certainly of relevance to my own study, of nineteenth-century pauper letters (Timmis 2020)). The precise question Biber and Egbert investigate is '[are there] written registers on the searchable web, readily accessible to researchers and practitioners, which provide reasonable representations of the typical discourse style found in spoken interaction?' (p. 319). Their corpus consists of 44,000 documents extracted from the *Corpus of Web-based English* (GloWbE). The analysis they carry out is highly register-sensitive and leads to the conclusion that there is a continuum of orality, but no single register is a close match for speech, as the linguistic characteristics of register are crucially shaped by situational factors (p. 331):

> … although song lyrics, transcribed interviews, TV transcripts, and discussion forums are the most 'oral' of the registers found on the public searchable web, their situational characteristics differ in several key respects from both spoken conversation and (super) synchronous CMC. It turns out that that these situation differences correspond to systematic linguistic differences.

The hallmark of all these chapters is rigorous methodology and a bold originality which combine to encourage the reader to look at language in new ways and from different perspectives, a fitting tribute indeed to the work of Merja Kytö.

Reviewer's address:
2 Victoria Crescent
Leeds LS18 4PR
UK
i.timmis@leedsbeckett.ac.uk

Reference

Timmis, Ivor S. 2020. *The discourse of desperation: Late nineteenth and early twentieth century letters by paupers, prisoners and rogues*. London: Routledge.

(Received 11 October 2021)

doi:10.1017/S1360674322000077

Sofia Rüdiger and **Daria Dayter** (eds.), *Corpus approaches to social media* (Studies in Corpus Linguistics 98). Amsterdam and Philadelphia: John Benjamins, 2020. Pp. vi + 210. ISBN 9789027207944.

Reviewed by Mikko Laitinen, University of Eastern Finland

Corpus approaches to social media contains a collection of chapters based on the papers presented in a workshop at the 40th ICAME conference (International Computer Archive

of Modern and Medieval English) in Neuchâtel. The volume presents contributions that address the issues related to the research design of studies that use social media as their primary material. Specific focus is placed on the ethics and legality of such endeavors, the technical expertise required in data collection and analysis, and the suitability of corpus linguistics methods in the study of material from social media.

The volume consists of nine chapters, the last of which is a commentary by Claire Hardaker. As their primary data the chapters draw on various social media applications such as Facebook, Reddit, Twitter and WhatsApp. A noteworthy feature of the volume is a digital companion website containing supplementary material related to the chapters. The contents are divided into three sections, the first of which focuses on the use of corpus methods in the study of online Communities of Practice (CoP). The second section tackles language variation in short social media texts, and contains chapters related to data from as yet inadequately studied platforms such as WhatsApp, and also presents novel methodologies usable in the study of short user-generated materials, viz. those that are shorter than texts that are often used in the study of language variation. The third section expands the scope to the study of multimodality and visual elements. In her discussion, Hardaker brings the various approaches together and provides a broader context for the topics in the various chapters.

The introduction defines the limits of the scope of the volume, but the editors also present a range of future desiderata for corpus-based research on social media. Sofia Rüdiger and Daria Dayter call for more interdisciplinary research and improved data access that will facilitate future work on social media by using corpus linguistic methods. In addition, their discussion highlights future crowdsourcing initiatives so that researchers would have better access to under-studied data sources; it also emphasizes related ethical issues and copyright questions.

In the first chapter, 'Towards a digital sociolinguistics: Communities of practice on Reddit', Sven and Martin Leuckert explore the extent to which the notion of a CoP can be applied to data from three subreddits. These 'are devoted to various topics and consist of discussion trees' (p. 19), and they could be expected to bring people together in mutual engagement in an endeavor. The authors look into one community related to professions (R/LINGUISTICS), another connected with e-games (R/LEAGUEOFLEGENDS), and a third with particular interest in LGBTQI+ rights (R/RUPAULSDRAGRACE). The contribution showcases how corpus methods can be utilized in sociolinguistic approaches to social media sources. The authors use crawled data that are subjected to frequency-based analysis and a similarity measure based on posting history. The results suggest that the concept of a CoP can be problematic in online settings, but corpus linguistics methodology turns out to be crucial for an understanding of the sociolinguistics in social media, such as how people employ the means of explicit self-identification or how they use shared in-group repertoires.

Lisa Donlan continues with the theme of CoP in 'The control and censorship of linguistic resources in an online Community of Practice' and looks into the ways in which community members engage in normative policing of linguistic choices in a

https://doi.org/10.1017/S1360674322000077 Published online by Cambridge University Press

subreddit of music enthusiasts (R/POPHEADS). The members engage in discussing, recommending and reviewing songs and albums, and as data points, they can be easily characterized in terms of power and may be defined by the length of their individual community membership, activity profiles and engagement. The empirical analysis focuses on one lexical item (*wig*) in the (modern slang) sense of being surprised or impressed (p. 47). The chapter presents a short-term diachronic study around events in which some members called for banning of the item. This led to a very slight decrease in the frequency of *wig*, but also substantial violations of the ban, often by members with significant power status in the community. Donlan concludes that there is a range of power types in a community, and knowing these predictors is crucial in understanding the multidimensionality of online power.

Dayter and Rüdiger in chapter 3 ('Talking about women: Elicitation, manual tagging, and semantic tagging in a study of pick-up artists' referential strategies') investigate how women are represented and discussed in a pick-up artist community. In short, pick-up artists aim at 'speed seduction of women' (p. 66), and the community relies on both online and offline communication. The chapter uses corpus-based discourse-analytic methods to study this community through a linguistic lens and makes use of introspection, manual tagging and reverse collocation search of semantic tags as its methods. The authors ask what type of referential strategies pick-up artists use to discuss women, and the results show that, while overtly derogatory terms are rarely used, there is a tendency to rely on infantilizing female referents in the discourse. The analysis also shows that the manual tagging, despite being time consuming, resulted in the highest accuracy, but the authors recommend using it alongside the more time-efficient elicitation and semantic tagging methods.

In part II the focus shifts to variation in short texts, and Samuel Felder looks into a corpus of WhatsApp messages primarily written in Swiss German in chapter 4 ('Patterns of intra-individual variation in a Swiss WhatsApp corpus'). The data were collected in 2014 and consist of about 750,000 messages extracted from dyadic chats. Felder investigates real-time change and long-term accommodation, looking at the use of punctuation, lexis (e.g. forms of *ja* and *jo* for 'yes'), and the use of emoji. The observations show that individual behavior in social media may change over the course of very short periods of time, and Felder accounts for these changes in terms of long-term accommodation in dyadic chats. The chapter identifies a range of processes of intra-individual variation, such as parallelism in which two individuals mutually influence each other in communication.

Aatu Liimatta (chapter 5, 'Using lengthwise scaling to compare feature frequencies across text lengths on Reddit') tackles the question of the comparability of texts of different lengths. He shows that various normalization methods reveal obvious complications when applied to genres that contain social media texts of highly variable length, such as Reddit posts. To illustrate: in the data used by Liimatta 50 percent of the comments are <20 words in length and 90 percent are under 100 words, and standard methods, like normalization, do not work in such settings. He pilots a method called lengthwise scaling, consisting of

https://doi.org/10.1017/S1360674322000077 Published online by Cambridge University Press

rarity scaling and quantile scaling, which are applied to each text length separately and which reduce the disparity between lengths and therefore enable more accurate comparisons across texts. The lengthwise approaches require large datasets and so Liimatta uses material from three Reddit communities with distinct genre profiles. The methods are illustrated using the first-person singular pronoun frequencies occurring in these communities, and the author concludes that the two lengthwise methods have obvious benefits, but he recommends quantile scaling as a particularly useful method.

In chapter 6, 'Double trouble: Are 280-character tweets comparable to 140-character tweets?', Martin Eberl tackles the question of what happened to language use in Twitter when the maximum length of tweets was increased from 140 characters to 280 in November 2017. He presents a longitudinal study that uses a dataset of messages prior to the increase and compares it with the post-switch situation. The data consist of material from almost 50,000 user accounts. The linguistic elements investigated consist of abbreviations (e.g. *'d* instead of *would*), type–token ratios, the use of punctuation, and logograms. The result is far from being a simple increase or decrease in the occurrence of certain forms, but Eberl shows that the increase in the tweet length had an impact, although his observations also show that tweets of varying length behave differently, something which needs to be taken into account in future studies. The author convincingly concludes that any future research that uses Twitter data ought to take the switch into account as a possible constraint for language use.

Part III shifts the focus to images, and chapter 7, 'Constructing corpora from images and text: An introduction to Visual Constituent Analysis' by Alex Christiansen, William Dance and Alexander Wild, uses visual constituent analysis, which makes it possible to explore visual aspects of online media through machine learning. Their approach treats images 'as a series of generic keywords or terms' (p. 150), and this enables the analysis of images as text that can then be further combined with textual material in online communication. They demonstrate the method using a sizeable dataset from state-backed information operations by the Russian Internet Research Agency, which was released by Twitter a few years ago. The empirical part investigates how images that have what the authors call an in-text reference differ from those with no textual material. The observations identify hidden discourses so that the two datasets show statistically different patterns in the semantic categories, such as the use of political slogans and specific hashtags.

Luke C. Collins, in chapter 8, discusses how images and emoji can be integrated into a corpus-based analysis. With due permission from the shop owners, he analyzes content from the Facebook pages of a local shop in Nottingham and uses a dataset of Facebook posts from one calendar year. The analysis excludes followers and customer-generated data. The dataset contains 1,167 images that were manually coded using an annotation scheme that enables images to be included in corpus searches. The contribution is a practical demonstration of data preparation but extends beyond that to showcase how multimodal elements can be included in wordlists and collocation analyses. The results show not only that including image content has an important function in social media

but also that image content might lead to rethinking core concepts in corpus linguistics, such as collocations.

In the final chapter Claire Hardaker provides a broader contextualization of the contributions. Her overview centers on major communicative events in human history, and she shows how immensely short is the period that the internet and social media have been in existence. She acknowledges that the work on social media has only just begun and cautions that the lack of chronological depth of data might easily lead researchers to erroneous conclusions. But she also illustrates how the contributions in this volume have offered first glimpses of potential new directions in the field. Her discussion aptly highlights the fact that, even though elegant software and algorithms now perform analyses far beyond anything that was possible only a decade ago, very little has actually changed in the lives of those whose language use is being analyzed, and that researchers should not lose sight of the human component in communication.

This is an interesting collection in which the contributions approach social media from a variety of perspectives. Obviously, the approaches are not intended to be comprehensive, but many of the contributions nevertheless add to current knowledge and highlight the fact that the tools and methods used in corpus linguistics are useful for analyzing social media. The three parts each contain a clear thematic focus, with the first of them being close to the traditional computer-mediated discourse analysis, in which the studies are complemented with quantitative corpus information. The second, in turn, is by far the most advanced technically and methodologically, while the third combines multimodal methods with corpus linguistics. An impatient reader may be left wondering about the slight methodological imbalance in these approaches, as some of the contributions clearly adopt more computational and algorithmic methodologies (e.g. Eberl; Liimatta), while others are more oriented towards the CoP framework and therefore sit closer to the discourse-analytic end of corpus linguistics. This can be a benefit, showcasing the breath of studies in this field, but it may also be a disadvantage, since the main theme of the volume may be lost.

A noteworthy positive feature in the volume is that nearly all of the visualizations are in color, which is not a default option with many publishers, who are generally reluctant to publish in color or charge outrageous sums for color images. We do not know if this has been the case here, but John Benjamins at least seems to be in the forefront, enabling the use of color images that are often required when visualizing large and rich datasets. On the negative side, the volume would have benefited from more careful editing, although this is a small issue. To illustrate, the chapter numbers mentioned in the introduction do not match the chapter numbering in the table of contents, and the significant contributions in using social media have already been presented in a special issue devoted to computational sociolinguistics rather than in 'computational dialectology' (p. 3).

Overall, however, the volume is of good quality and presents a promising area for corpus linguistics. In the introduction, the editors present a vision for the future by highlighting issues such as research ethics and data availability. They suggest that data access is a frequently encountered problem in social media research. Given this statement, with which the present reader fully agrees, it is surprising that the primary data used in the studies are not available, for instance, in the digital appendix. Some studies (Leuckert & Leuckert) do

contain links to the scripts used, but to ensure reproducibility and replicability it might be useful to develop shared practices in the field. English corpus linguistics, with its long tradition of making data available, might be at the forefront of introducing such practices.

When moving beyond the contents of this individual volume to the broader theme of using social media data in English linguistics, it is clear that social media data have great potential in the evidence-based study of English (including corpus linguistics, sociolinguistics, and language variation and change). I would suggest, however, that social media data have characteristics that call for completely novel approaches, and relying on a single set of, albeit solid, methods provides insufficient results and an incomplete picture of the phenomena under study. The editors of this volume also point out, somewhat modestly, that there is an increasing need to engage in more interdisciplinary research in the study of social media. A critical reader might argue that, to understand and fully benefit from very large and often rich social media in corpus linguistics in the future, a mere integration of quantitative and qualitative alone is insufficient. The sheer size and complexity of data (both user-generated textual data and metadata) present us with a research setting that calls for transdisciplinary approaches, and also highlights the need to broaden the expertise in computational and algorithmic directions. Corpus linguists, with long traditions in combining methods, are ideally positioned to engage in fuller collaboration with, for instance, researchers in AI, computational linguists, visualization experts and data mining specialists. The present volume is a good start in that direction, but we still need a fuller integration of methods and competencies in the future.

Reviewer's address:
School of Humanities, Foreign Languages and Translation Studies
University of Eastern Finland
PO Box 111
FI-80101 Joensuu
Finland
mikko.laitinen@uef.fi

(Received 7 April 2022)

doi:10.1017/S1360674322000260

Merja Stenroos and **Kjetil V. Thengs** (eds.), *Records of real people: Linguistic variation in Middle English local documents* (Advances in Historical Sociolinguistics 11). Amsterdam: John Benjamins, 2020. Pp. viii + 310. ISBN 9789027207951.

Reviewed by Jacob Thaisen, University of Oslo

Part I of *Records of real people: Linguistic variation in Middle English local documents* lays out the Middle English Scribal Text (MEST) programme's theoretical stance. It