Corpus Linguistics and Social Media

William Dance, Department for Linguistics and English Language, Lancaster University, Lancaster, United Kingdom

Key Points

- Social media platforms offer researchers vast amounts of data suitable for large-scale linguistic analysis.
- Corpus linguistics, with its emphasis on naturally occurring datasets, is well-suited to the study of social media, its structure and understanding how ideas spread online.
- Understanding the social media discourses surrounding key political, economic and socio-cultural topics can provide key insights other methods often overlook.
- The analysis of social media data allows us to update, and challenge, foundational understandings in studies of communication and linguistics.
- Social media platforms are constantly changing, and corpus linguistics can leverage these new developments to help us understand important socio-technological issues online.

Abstract

Social media services are used by billions worldwide to share user-generated content. This content offers a vast array of insights to linguists interested in structure, meaning, interaction and other elements of communication. The size of the datasets, however, can overwhelm many approaches to language. Corpus linguistic tools that use computer-aided methods to analyze large datasets are well-suited to these data sources. The study of social media using corpus approaches allows us to understand how social media has changed how we communicate, and how key ideas are mediated through language online.

Introduction

Social media services such as Facebook, X, and TikTok offer a wealth of user-generated content for researchers to explore. These services are accessed by billions worldwide and act as key information exchanges for many individuals, businesses, and organisations. Corpus linguistics, the analysis of large, naturally occurring datasets (McEnery & Hardie, 2011), is well-suited to the study of social media and offers a way to study large amounts of social media data with due nuance.

Corpora can be used to answer research questions and test hypotheses on real-world data and for over a quarter of a century, the World Wide Web (hereafter "internet") has been regarded as the largest corpus ever developed:

> "[The internet is] the biggest searchable language corpus we have ever possessed, and its undoubted usefulness in linguistic research should be exploited […] we have at our disposal the biggest corpus of electronic texts so far possessed—a supercorpus, as it were!"
>
> Bergh et al. (1998), pp. 47, 53.

The "supercorpus" that Bergh et al. (1998) refer to is estimated to be around 8-billion words in size. This impressive size, however, pales in comparison to the size of modern-day social media services. While official statistics can be elusive, it is estimated that daily 4.75 billion items are shared to Facebook (Bagadiya, 2024), 500 million tweets are sent (X, 2014), and 24 million videos are posted to TikTok (Daniel, 2024). This means that in less than a day, enough content is produced on social media to dwarf the

corpus estimation by Bergh et al. (1998). Ever since, the internet and social media services have grown exponentially in size and offer massive amounts of naturally occurring, machine readable data for corpus linguists.

Corpus Linguistics and the Participatory Web

The first iteration of the internet, Web 1.0, comprised mostly one-way information consumption where users were largely consumers of information and not producers. Web 2.0, known as the participatory web, is characterized by users generating and sharing content with each other. The transition to Web 2.0 roughly coincided with the turn of the millennium and the earliest examples of the participatory web include websites such as message boards and forums. These services allowed the exchange of user-generated content with other internet users. Herring (2013, p. 4) defines Web 2.0 as:

> Web-based platforms that emerged as popular in the first decade of the 21st century, and that incorporate user-generated content and social interaction, often alongside or in response to structures and/or (multimedia) content provided by the sites themselves.

Websites associated with Web 2.0 such as internet forums offered a wealth of naturally occurring (i.e., non-elicited) linguistic data and thus were very useful to corpus linguists. Early uses of the social web as a corpus includes the Hungarian National Corpus (HNC) which contained 20-million words of internet forum data (Váradi, 2002), and the Net-EN corpus (NC) containing 546,000 tokens of forum data (Takahashi, 2003). By the 2010s, vast social media services had formed across the world, with platforms such as Facebook, X (then "Twitter"), Weibo, VKontakte (VK), and others generating massive amounts of data from billions of users. **Table 1** shows all the social media companies globally with more than 100 million monthly active users (MAUs) (Statista, 2023).

At present, over 20 social media platforms have more than 100-million monthly active users and these ever-growing data sources allow researchers to explore various aspects of language, including grammar, interaction, and the differences between written and spoken language. Studies of social media are related closely to the field of computer-mediated communication and the study of online communication that more broadly apply to digital communication in general, extending beyond just social media and social networking platforms.

Accessing and Processing Social Media Corpora

While social media services are already in a machine readable format, they are not necessarily readily available for analysis. Access to social media data varies and while some services are openly available, others require users to log in to view content. This can confound automated data collection methods. Some social media services offer an Application Programming Interface (hereafter "API"), an official data stream that allows people to access data from the service. Up until 2023, X provided an Academic Access API that allowed free access to 10,000,000 tweets a month, a service that has since been replaced by a fee-paying option. Elsewhere, TikTok offers academic research access and Reddit can be accessed through the Pushshift API (Baumgartner et al., 2020). Other social media services are less receptive to research, with services such as Facebook offering virtually no official access to researchers.

Social media posts also extend beyond just the content of the post itself. Information about the time something was posted, where it was posted from, and other details, can also be used for analysis. This data about data, or "metadata", can be added into corpora using Extensible Markup Language (XML), a way of encoding additional information into data.

The tweet in **Fig. 1** collected using X's Academic Access API in 2021 shows XML in action. The textual data, i.e., the tweet itself, is shown in black, while additional metadata about the tweet is shown in red and purple. This metadata shows that the tweet originates from the BBC News verified technology news account, who disclose their location as "Glasgow, Scotland" and had 907,471 followers at the time of collection. This additional information can be used in various ways, such as to create sub-corpora, isolate specific temporal trends, or investigate sociolinguistic research questions using demographic data.

Table 1 Monthly actives users (MAUs) on social media platforms with more than 100 million users worldwide (Statista, 2023).

Rank	*Platform*	*MAUs (billions)*	*Rank*	*Platform*	*MAUs (billions)*	*Rank*	*Platform*	*MAUs (billions)*
1	Facebook	3.05	8	Snapchat	0.75	15	Reddit	0.43
2	WhatsApp	2.78	9	Kuaishou	0.673	16	LinkedIn	0.424
3	YouTube	2.49	10	Weibo	0.599	17	Quora	0.3
4	Instagram	2.04	11	QQ	0.571	18	Discord	0.154
5	WeChat	1.32	12	Qzone	0.558	19	Twitch	0.14
6	TikTok	1.22	13	X (Twitter)	0.55	20	Tumble	0.135
7	Telegram	0.8	14	Pinterest	0.465	21	Threads	0.1

```
<tweet> Google, Youtube ban ads on climate misinformation https://t.co/N8Q5L1Btde
</tweet> < id='1446480230491856900' createdAt='2021-10-08T14:18:55.000Z' language
='en' authorId='621583' authorUsername='BBCTech' authorName='BBC News Technology'
 authorVerified='TRUE' authorDescription='The official account for the BBC News
technology team: @scottishmon @zsk @jamesclayton5 @shionamc' authorLocation=
'Glasgow, Scotland' authorCreatedAt='2007-01-10T12:41:22.000Z'
authorFollowersCount='907471' authorFollowingCount='65' authorTweetCount='34397'
authorListedCount='15498' referencedTweetId='' referencedTweetCreatedAt='' >
```

Fig. 1 An example of metadata for a tweet.

The collection of social media posts, and accompanying user information, inevitably raises the topic of ethics. At the core of this issue is whether social media posts constitute public data as the distinction between public and private online is very much blurred. The British Sociological Association (2016) (BSA) provide a comprehensive overview of this topic, giving specific attention to the distinction between private and public communication and what this means for research. There is also considerable variation both between social media services and within them meaning a one-size-fits-all approach does not work, leading to the need for a "situational ethics" approach that prioritizes confidentiality while allowing for a pragmatic approach that allows for "discretion, flexibility, and innovation" (British Sociological Association, 2016, p. 16).

Social Media Discourse

Social media can be analyzed both as a text type to identify the characteristics of a given social network, and as a medium through which important topics are discussed. Considering what makes social media language unique, for example, Zappavigna (2011) (2015) explores the use of hashtags "#" on what was then Twitter. Zappavigna finds that beyond their core function of indexing tweets for others to find, hashtags function as a key semiotic resource by "marking experiential topics, enacting interpersonal relationships, and organizing text" at the grammatical, semantic, and discourse level (Zappavigna, 2015, p. 274). Elsewhere, McCulloch and Gawne (2018) propose that emoji (small pictures used as text) should be viewed as digital gestures as they lack their own grammar, but still contribute to meaning through repetition and emblematic meaning. These studies show how we can use social media to explore structure, meaning, and interaction.

Many studies have investigated social media language to understand how topics are represented online and to identify and examine discourses, the sets of meanings and representations that surround a topic (Burr, 1995). Analyzing social media discourse involves carrying out what Herring (2004) calls Computer-Mediated Discourse Analysis (CMDA), the adaptation of language-focused disciplines to study social media data:

> Indeed, the potential—and power—of CMDA is that it enables questions of broad social and psychological significance, including notions that would otherwise be intractable to empirical analysis, to be investigated with fine-grained empirical rigor.
>
> Herring (2004), p. 340.

Social media platforms are used for a wide range of purposes, from political participation through to simply keeping up with friends. They offer users an online space to engage others in conversation, debate and general information sharing, and allow individuals to reproduce, or reinvent, their offline identities (Hardaker & McGlashan, 2016). This sometimes unfiltered participation means social media services have acted as an invaluable resource for understanding how important issues are mediated through language. These studies can have implications for our understandings of key issues across a range of topics and complement existing methods such as surveys, questionnaires, and interviews. One of these key areas is health.

In a study of the parenting forum Mumsnet Talk, Semino et al. (2023) explore resistance to persuasion by using corpus-based discourse analysis of health discussions. The extant literature in this field relies mostly on experimental studies where data is elicited from participants. By choosing to analyze naturally occurring, organically evolving data in discussions of HPV vaccination however, their methodological innovation allowed them to further develop existing typologies of resistance to persuasion using real-world data. This builds on other research that uses corpus methods on social media data to understand vaccine hesitancy (Coltman-Patel et al., 2022). Such studies are valuable for policymakers and healthcare professionals and show the potential of using social media corpora.

Social media data has also proved useful in updating, and sometimes challenging, key theories of language and discourse. In relation to changing our understandings of critical discourse studies (CDS) in the era of social media, KhosraviNik (2017) describes how social media has fundamentally challenged our "traditional static understanding of media power" which has implications for how we do research in digital discursive environments. KhosraviNik suggests that the participatory nature of social media interactions, and the involvement of users in the creation of texts, has changed the structural norms of communication and this has eroded the power structures behind discourse. This change in how discourse is generated, and disseminated, has meant a departure from the mainstream media model, and has resulted in the increased empowerment of the individual. This changes how we understand the

critical, socially situated study of language. These shifts not only have theoretical impacts on our understanding of critical discourse studies, power and ideology, but also methodological implications for how we understand texts that are socially situated into varying, evolving technological practices.

Audio, Images, and Videos

Social media services have become increasingly multimodal. Not only are features such as GIFs, stickers, and emojis commonplace (Collins, 2019), but social media services are also designed with multimedia in mind. In 2013, Facebook announced the service would become more photo-centric (Facebook, cited by Robertson, 2013), while in 2023, X revealed that tweets with images are boosted and shown to double the audience compared to purely textual tweets (X, 2023). Further, with the rise of short-form video sharing services such as TikTok and Reels, relying on purely textual social media data is becoming outdated.

Multimodal social media data offers both possibilities and complications. In practical terms, audio and video files take up much more space than text files and therefore require greater storage capabilities. They also require additional processing to be used in a corpus processor. Audio and video data must first be transcribed, and decisions must be made on how to annotate image data to make it machine-readable. These processes require additional time, resources, and costs.

There have been many recent developments in this field. Services such as Mechanical Turk (MTurk) allow researchers to crowd-source image annotation from the public, a process known as microtasking (Gadiraju et al., 2015). Although it alleviates some issues such as time, microtasking increases cost and additionally requires a coding scheme to be developed for the microtaskers. Elsewhere, some studies use automated image annotation tools such as Google's Cloud Vision to translate visual data into textual data that can be processed by traditional corpus processors (Baker & Collins, 2023; Christiansen et al., 2020). These computer vision tools identify features in images to convert them from visual to textual data. **Fig. 2** shows some labels ascribed by Google Cloud Vision to an image, these include "tree" and "agriculture" for a photo of a wheat field.

The Validity of Social Media Data

Many argue that social media sentiments, opinions, and trends do not mirror those of offline attitudes and behaviors. The reason for this is often that social media represent a polarized, vocal minority and subsequently studies of social media are not truly representative of offline contexts. For example, the online disinhibition effect can lead to individuals expressing things online that they would not normally express offline (Suler, 2004), whether these are positively or negatively marked behaviors. However, rather than using a binary offline-online distinction, it is better instead to view the online disinhibition effect as a spectrum, ranging from certain highly anonymous contexts where speakers have total freedom, to contexts where individuals are enacting their digital selves and therefore act largely as they would offline.

Others also argue for the similarity of social media communication and face-to-face interaction, asserting that social media language should simply be seen as an adaption of oral speech, and not as a distinct form (Benwell & Stokoe, 2006; Meredith & Stokoe, 2013). Further, given the increasingly sophisticated integration of social media into our lives, through devices such as mobile phones, wearable technology (E.G. smartwatches), and virtual reality, this distinction between online and offline is becoming increasingly indistinct. In other words, what is online and offline will gradually mean less over time.

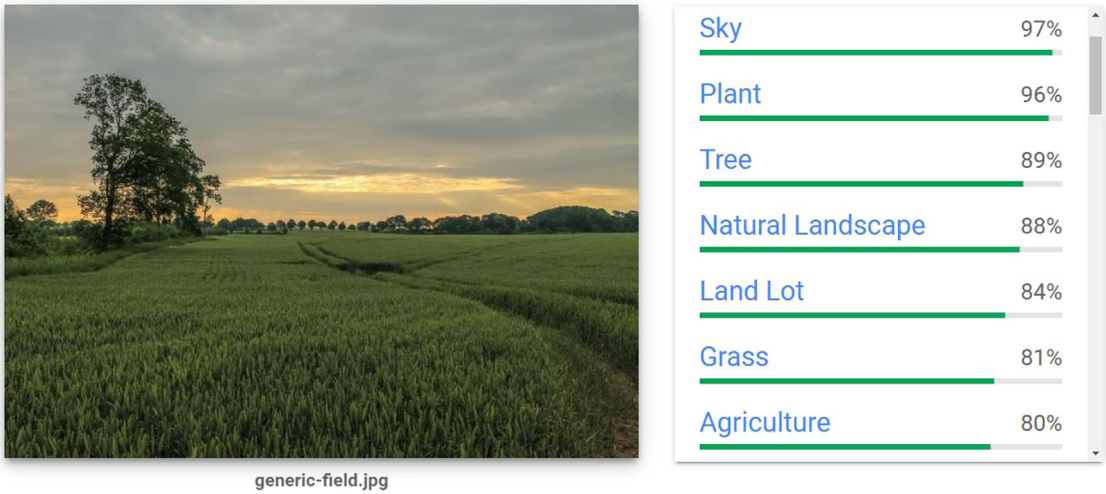

Fig. 2 An example of google cloud vision image tagging output (Stockvault).

The increasing availability and accessibility of generative artificial intelligence (AI), tools that create text, images, audio, and videos as if they were human produced, also affects corpus linguistics. These tools have led to an increased proliferation of artificially produced language across social media platforms, challenging the *authentic* component of many textbook definitions of corpora (Bennett, 2010; McEnery et al., 2006). Alongside this, vast quantities of inauthentic accounts, known as "bots", can also artificially skew social media discussions, leading to corpora that may represent artificial language and not naturally occurring language. Other digital discursive practices such as zone flooding, where social media accounts intentionally overload the information environment to drown out authentic discourse (Dance, 2024), further compound these issues. Given the role of corpora in studying real-world language and answering research questions and testing hypotheses on real-world data, these considerations are important for corpus practitioners.

Conclusion

This article has shown how corpus linguistics can be used in the study of social media, how our understandings of large corpora have changed over time, and the utility of using corpus approaches to study social media. The study of social media is still a relatively young field and as a method and a discipline, corpus linguistics is well-suited to the study of social media texts and practices. Corpus linguistic tools enable the researcher to analyze vast quantities of social media data while also carrying out sentence-level qualitative analysis that approaches such as natural language processing (NLP) simply do not allow for. This makes corpus linguistics a valuable tool for understanding how social media changes language, and for understanding key topics mediated through social media discourses. Corpus linguistics will equally have to adapt to new technologies. As social media companies become more innovative with the increasing popularity and availability of features such as AI and virtual reality, corpus linguistics is poised to allow researchers to be at the center of key technological and social change.

References

Bagadiya, J. (2024). *38 Facebook statistics and facts for every marketer in 2024*. Retrieved from https://www.socialpilot.co/facebook-marketing/facebook-statistics.

Baker, P., & Collins, L. (2023). Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's automatic image tagger. *Applied Corpus Linguistics, 3*(1), Article 100043. https://doi.org/10.1016/j.acorp.2023.100043

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. In *Paper presented at the proceedings of the international AAAI conference on web and social media*.

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor, Mich: University of Michigan Press.

Benwell, B., & Stokoe, E. (2006). *Discourse and identity*. Edinburgh University Press.

Bergh, G., Seppänen, A., & Trotta, J. (1998). Language corpora and the internet: A joint linguistic resource. In *Explorations in corpus linguistics* (pp. 41—54). Brill.

British Sociological Association. (2016). *Ethics guidelines and collated resources for digital research*. Retrieved from https://www.britsoc.co.uk/media/24309/bsa_statement_of_ethical_practice_annexe.pdf.

Burr, V. (1995). *An introduction to social constructionism* (1 ed.). Routledge.

Christiansen, A., Dance, W., & Wild, A. (2020). Constructing corpora from images and text. *Corpus Approaches to Social Media*, 149—174.

Collins, L. (2019). *Corpus linguistics for online communication: A guide for research*. Routledge.

Coltman-Patel, T., Dance, W., Demjén, Z., Gatherer, D., Hardaker, C., & Semino, E. (2022). "Am I being unreasonable to vaccinate my kids against my ex's wishes?"—a corpus linguistic exploration of conflict in vaccination discussions on Mumsnet Talk's AIBU forum. *Discourse, Context & Media, 48*, Article 100624. https://doi.org/10.1016/j.dcm.2022.100624

Dance, W. (2025). Disinformation and algorithms: Amplification, reception and correction. In S. Rüdiger, & D. Dayter (Eds.), *Manipulation, influence and deception: The changing landscape of persuasive language*. Cambridge University Press, 9781009098724.

Daniel, C. (2024). *TikTok users and growth statistics*. Retrieved from https://www.usesignhouse.com/blog/tiktok-stats.

Gadiraju, U., Demartini, G., Kawase, R., & Dietze, S. (2015). Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems, 30*(4), 81—85.

Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics, 91*, 80—93. https://doi.org/10.1016/j.pragma.2015.11.005

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. *Designing for Virtual Communities in the Service of Learning, 338*, 376.

Herring, S. C. (2013). Discourse in web 2.0: Familiar, reconfigured, and emergent. *Discourse, 2*(0), 1—26.

KhosraviNik, M. (2017). Social media critical discourse studies (SM-CDS). In *The Routledge handbook of critical discourse studies* (pp. 582—596). Routledge.

McCulloch, G., & Gawne, L. (2018). *Emoji grammar as beat gestures. Paper presented at the Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media, Standford [en línea]*. Disponible en http://knoesis.org/resources/Emoji2018/Emoji2018_Papers/Paper13_Emoji2018. pdf. (Accessed 11 December 2019).

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Meredith, J., & Stokoe, E. (2013). Repair: Comparing Facebook "chat" with spoken interaction. *Discourse & Communication, 8*(2), 181—207. https://doi.org/10.1177/1750481313510815

Robertson, A. (2013). *Facebook redesigns News Feed with multiple feeds and "mobile-inspired" interface*. Retrieved from https://www.theverge.com/2013/3/7/4075548/facebook-redesigns-news-feed-with-multiple-feeds.

Semino, E., Coltman-Patel, T., Dance, W., Deignan, A., Demjén, Z., Hardaker, C., et al. (2023). Narratives, information and manifestations of resistance to persuasion in online discussions of HPV vaccination. *Health Communication*, 1—12.

Statista. (2023). *Most popular social networks worldwide as of October 2023, ranked by number of monthly active users*. Retrieved from https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior, 7*(3), 321—326.

Takahashi, J. A. (2003). Do we talk (or write?) differently over the Net? A lexical enquiry into"a"Net-EN. In *Paper presented at the proceedings of the corpus linguistics 2003 conference (CL2003)*. UK: Lancaster, Lancaster University.
Váradi, T. (2002). The Hungarian national corpus. In *Paper presented at the LREC*.
X. (2014). *The 2014 #YearOnTwitter*. Retrieved from https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html.
X. (2023). *Twitter/the-Algorithm: Source code for twitter's recommendation algorithm*. Retrieved from https://github.com/twitter/the-algorithm.
Zappavigna, M. (2011). Ambient affiliation: A linguistic perspective on twitter. *New Media & Society, 13*(5), 788–806.
Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social Semiotics, 25*(3), 274–291. https://doi.org/10.1080/10350330.2014.996948