Journal of Al-Farabi for Humanity Sciences Volume (8), Issue (6) December (2025)

ISSN: 2957-3874 (Print)
Journal of Al-Farabi for Humanity Sciences (JFHS)
https://iasj.rdd.edu.iq/journals/journal/view/95
مجلة الفارابي للعلوم الإنسانية تصدرها جامعة الفارابي

Hate Speech vs. Free Speech: A Pragmatic Approach to Boundary- Marking in Online Debates

Asst. Lect. Balsam Yousif Yaqoub
General Directorate of Education in Wasit

خطاب الكراهية مقابل حرية التعبير: مقاربة براغماتية لتحديد الحدود في النقاشات الإلكترونية

مساعد محاضر: بلسم يوسف يعقوب
المديرية العامة للتربية في واسط

Abstract:
With the advent of the internet, the conflict between free speech and hate speech has been further aggravated, especially on the internet where the boundaries between harmless and harmful communication are continually blurred. The research delves into how online discourse participants pragmatically manage the controversial boundaries between free speech and hate speech. The research problem is the ambiguity and inconsistency in operationalization, interpretation, and definition of these two notions in online communication, which provokes polarized disagreement and argumentation. The main aim is to investigate the linguistic and pragmatic mechanisms through which users defend, challenge, or legitimize protected or offending language. Data were garnered from Twitter and Reddit comment threads, and the threads were those related to contentious topics like immigration, gender identity, and political dissent. An eclectic analysis model was utilized, which borrowed from speech act theory, Gricean pragmatics, and aspects of critical discourse analysis in examining both micro-linguistic choice and macro-discursive formations. Findings show that participants frame their messages with appeals to intentions, rights, and social norms in an attempt to cast themselves as defenders of free speech or victims of hate speech. Further, site-specific norms and contextual markers are crucial in the interpretation and contestation of speech. The research calls for a sophisticated, context-sensitive appreciation of speech boundaries and argues that pragmatic sensitivity can help separate hate speech from acceptable discourse. The findings have implications for moderation policy, online literacies, and the law governing online communication.Keywords: hate speech, free speech, pragmatics, online discourse, boundary-marking, speech act theory.

الملخص

مع ظهور الإنترنت، تزايد الصراع بين حرية التعبير وخطاب الكراهية، لا سيما على الإنترنت حيث تتلاشى الحدود بين التواصل غير الضار والتواصل الضار بشكل مستمر. تتناول هذه الدراسة كيفية إدارة المشاركين في النقاشات الإلكترونية للحدود المثيرة للجدل بين حرية التعبير وخطاب الكراهية من منظور براغماتي. تتمثل مشكلة البحث في الغموض وعدم الاتساق في تفعيل وتفسير وتعريف هذين المفهومين في التواصل عبر الإنترنت، الأمر الذي يؤدي إلى انقسام الخلافات وظهور الجدل الحاد .يهدف البحث إلى دراسة الآليات اللغوية والبراغماتية التي يستخدمها المستخدمون للدفاع عن اللغة المحمية، أو تحديها، أو تبريرها في حال اعتُبرت مسيئة. تم جمع البيانات من سلاسل التعليقات على منصتي تويتر وReddit، مع التركيز على المواضيع الجدلية مثل الهجرة، والهوية الجندرية، والمعارضة السياسية. اعتمد البحث نموذج تحليل انتقائي يجمع بين نظرية الأفعال الكلامية، والبراغماتية الجريسية، وبعض جوانب تحليل الخطاب النقدي، لفحص الاختيارات اللغوية الدقيقة (الميكرو–لغوية) والتكوينات الخطابية الكبرى (الماكرو–خطابية) .أظهرت النتائج أن المشاركين يهيئون رسائلهم بالاعتماد على الإشارات إلى النوايا، والحقوق، والمعايير الاجتماعية في محاولة لتقديم أنفسهم كمدافعين عن حرية التعبير أو كضحايا لخطاب الكراهية. علاوة على ذلك، تلعب المعايير الخاصة بالمواقع والمؤشرات السياقية دورًا جوهريًا في تفسير الخطاب ومناقشته. يدعو البحث إلى تقدير متقدم وحساس للسياق لحدود الخطاب، ويؤكد أن الحساسية البراغماتية يمكن أن تساعد في التمييز بين خطاب الكراهية والخطاب المقبول. تحمل النتائج دلالات مهمة على سياسات الرقابة، ومهارات التعامل

مع المحتوى الرقمي، والقوانين المنظمة للتواصل عبر الإنترنت.الكلمات المفتاحية :خطاب الكراهية، حرية التعبير، البراغماتية، النقاش الإلكتروني، تحديد الحدود، نظرية الأفعال الكلامية.

1. Introduction

The rise of online communication has profoundly transformed how individuals participate in public discourse. Social media platforms, comment sections, blogs, and online discussion forums have emerged as central locations for formulating opinions, contesting mainstream discourses, and debating sociopolitical questions. Though such spaces have expanded public discourse inclusivity and reach, they have also amplified contention regarding the boundaries of acceptable speech. Specifically, the conflict between hate speech and free speech has become a focal point of debate in recent online discourse. As people claim their freedom of speech, others also appeal to the necessity for protection against harmful, discriminatory, or violent inciting speech. This double nature raises numerous challenges, both legal, ethical, and practical—particularly in contexts where regulation is minimal and understandings of significance are highly context-dependent.The central issue of concern in this research is the uncertainty regarding the definitions, meanings, and laws of hate speech and free speech in online communication. Although legal codes in most nations strive to create boundaries between protected and prohibited speech, such demarcations frequently find themselves lacking when confronted with the ephemerality and informality of online communication. In online environments such as Twitter and Reddit, characterized by brief, unplanned, and often decontextualized messages, users have to cope with intricate communicative ecologies. Such fluidity enables individuals to construct contentious speech acts in manners that have the effect of concealing or legitimizing their possible negative consequences. Additionally, understanding such discourse relies on cultural norms, ideological alignments, and norms particular to a certain platform, making boundary-drawing a highly pragmatic and context-dependent undertaking (Brown & Levinson, 1987; Haugh, 2013).The objectives of this study are threefold. First, it seeks to explore the linguistic and pragmatic devices through which online contributors depict their speech acts as either authentic expressions of opinion or offending instances of hate speech. Second, it strives to explore the contextual and interactive features that give rise to their interpretation. Third, it examines how these practical negotiations instantiate and constitute broader ideological positions, and influence digital norms and moderation policies. To achieve such ends, the analysis adopts an eclectic model of analysis drawing on speech act theory, Gricean pragmatics, and critical discourse analysis. This multi-dimensional approach allows for a subtle examination of both the micro-level (e.g., lexical selection, illocutionary force) and macro-level (e.g., power dynamics, ideological stance) features of online discourse (Searle, 1969; Grice, 1975; Fairclough, 1995).The following research questions underpin the study:

1. What pragmatic means are employed by users to define the boundary between hate speech and free speech in online discourse?
2. In what ways do contextual cues and platform rules influence the meaning of contested speech acts?
3. How do participants ideologically position themselves in online interactions through speech framing?

It is worth knowing how individuals pragmatically traverse these boundaries for several reasons. First, it tells us about the real-time linguistic and social mechanisms by which speaking norms are created and contested online. Second, this knowledge is crucial for informing effective and respect-oriented platform moderation policies. Third, it is part of the broader discussion of digital literacy, allowing users to be more attuned to how language operates online. Lastly, it has implications for democratic societies' legal and ethical frameworks seeking to manage online speech, where the balance between protection and expression is an ongoing challenge (Waldron, 2012; Butler, 1997).Generally, this research addresses an emergent and urgent communicative challenge through an examination of the pragmatic processes through which speech boundaries are negotiated in online spaces. In doing so, it highlights the need for more flexible and context-aware understanding of language working in the digital public sphere.

2. Literature Review The free speech and hate speech controversy has produced a vast interdisciplinary body of scholarship, from linguistics, philosophy, and law to communication studies, media studies, and digital media research. Such scholarship primarily examines the conceptual boundaries of these two species of speech and the socio-political and legal implications of either. Pragmatics offers a helpful lens on how meaning is created, negotiated, and fought over in face-to-face interaction—especially in online situations.Legally and philosophically, free speech has been considered a pillar of democratic societies, codified in constitutions and international human rights instruments (Barendt, 2005; Mill, 1859/2008). The right is not absolute. Hate speech, traditionally recognized as speech that incites violence, hatred, or discrimination against persons or groups

based on attributes such as race, religion, gender, or sexual orientation, is often deemed not to be protected by certain bodies of law (Schauer, 1982; Heinze, 2016). Jurists have struggled to strike a balance between protecting freedom of expression and preventing harm, with the need for balanced approaches considering context and potential consequences being emphasized (Gelber & Stone, 2007).In linguistics and pragmatics, researchers have examined the ways in which speech acts come into play where meanings are not necessarily implicit. Searle's (1969) speech act theory makes a distinction between the literal meaning of an utterance and its illocutionary force, illuminating how speakers employ malicious intent under the guise of free speech. Grice's (1975) conversation maxims and implicature theory also explain how implied meaning assists in understanding what appears to be offensive or harmful statements. They have been crucial in studying how participants manage face-threatening acts, politeness, and aggression in communication (Locher & Watts, 2005; Culpeper, 2011).Online spaces possess the uncertainty of intention and lack of paralinguistic cues that make it difficult to identify hate speech (Marwick & boyd, 2011). Online media increase the reach and visibility of user-created content that occasionally does not have the editorial check in mainstream media. As such, users will engage in metapragmatic practice—commenting on or positioning other people's speech—to contest or explain utterances (Blommaert, 2005). Several researches have investigated how online users justify offensive material by citing free speech, employing discursive tools such as irony, parody, or victimhood to reclaim themselves (Phillips, 2015; Graham & Hardaker, 2017).Critical discourse analysis (CDA) develops this area further by analyzing how power relations and ideologies are embedded in the use of language (Fairclough, 1995; van Dijk, 1998). It has been confirmed through CDA studies that hate speech is usually framed not as overt slurs but as "reasonable concerns" or "truth-telling," thus concealing its discriminatory nature while simultaneously claiming discursive legitimacy (Wodak, 2015). These inclinations are particularly pronounced on the extreme right and populism rhetoric, where racial and gender animus is masked through euphemistic or coded language (Krzyżanowski, 2020).Previous studies have also examined the ways in which platform affordances and norms shape discourse. For instance, differences in moderation practices on Twitter, Facebook, and Reddit influence what people say and how speech is supported or contested (Jhaver et al., 2019; Matamoros-Fernández, 2017). The phenomenon of "context collapse," where different audiences with disparate norms find themselves present in a single space, similarly complicates pragmatic interpretation (Marwick & boyd, 2011).While there is richness in existing literature, there remains a gap within research that unites pragmatic theory with real-time digital discourse analysis to try to analyze how users actively mark and negotiate these boundaries between free speech and hate speech. The present study seeks to address this gap using.an eclectic analytical model where speech act theory, Gricean pragmatics, and CDA meet to give a more contextually informed and dynamic view of online boundary-marking practice.

3. Methodology

3.1. Nature of the Study The qualitative methodology based on discourse pragmatics and critical discourse analysis (CDA) is employed in this study to investigate the ways online users navigate the contentious boundaries between hate speech and free speech. A qualitative method of inquiry is ideally suited to the capture of finely grained by context-dependent nature of linguistic meaning, particularly in user-created, on-the-fly textual material (Denzin & Lincoln, 2018). By focusing on user interpretation practices in real online conversation, the study identifies ways in which meaning is being co-constructed and negotiated within online public spheres.

3.2. Data Collection and Description The data was collected from two of the most widely used online media: Twitter and Reddit. The media were chosen due to their popularity, exposure to public discourse, and frequency at which political and contentious topics are discussed. Purposive sampling was employed in selecting comment threads on contentious topics such as immigration, LGBTQ+ rights, race issues, and political opposition. Gathering data focused on threads with high activity (i.e., over 100 comments or replies) and collated reported instances of flagged or complained-about speech.The last corpus had approximately 150 comment threads (75 Reddit and 75 Twitter), with a total of over 50,000 words. Date, platform, and post engagement (replies, likes, shares) were also recorded as metadata. All user handles were anonymized, and no personally identifiable information was retained to keep the participants anonymous.

3.3. Modal of the Study The ecological analytical framework involved an integrated use of three complementary approaches: speech act theory (Searle, 1969), Gricean pragmatics (Grice, 1975), and critical discourse analysis (Fairclough, 1995). Speech act theory enabled the classification of utterances based on their illocutionary force—whether assert, accuse, insult, or defend. Gricean pragmatics, particularly the theory of

implicature and conversational maxims, was used to explore how users imply or disavow hateful intent indirectly. Finally, CDA provided the tools for deconstruction of the way power, ideology, and identity were inscribed in and negotiated through such discourses (Wodak & Meyer, 2016).

3.4. Procedures of Analysis The analysis was conducted in four steps. In step one, the data were read multiple times with a view to identifying emerging themes of the contestation or justification of speech boundaries. In step two, utterances were manually coded with NVivo software using a developed coding scheme derived from speech functions (i.e., defending free speech, accusing of hate speech, ironic denial, ideological framing). Third, Gricean maxims of relevance, manner, and quantity were used to examine how pragmatic violations generated implicatures that revealed or hid speaker intentions (Thomas, 2013). Finally, through the CDA model, the study examined how such pragmatic choices realized themselves in broader ideological positions, such as appeals to liberalism, nationalism, or identity politics. At each step, careful consideration was placed on context: surrounding discourse, platform conventions, and inferred audience expectations.

3.5. Ethical Considerations As the research was on publicly available web conversation, it followed digital research ethics as established by the Association of Internet Researchers (AoIR, 2019). All accessed data were derived from public discussion platforms upon which users are necessarily provided with no guarantee of privacy. Although this, great care was taken to ensure that user anonymity was protected. Usernames and quotations were paraphrased wherever possible to ensure non-traceability. No content from private messages or closed forums was used. Additionally, platform terms of service were reviewed to ensure compliance. The study was reviewed and approved by the university institutional ethics committee prior to carrying out the data collection.

4. Data Analysis This section presents an analysis of six representative extracts from Reddit and Twitter debates, using the eclectic model combining Speech Act Theory, Gricean Pragmatics, and Critical Discourse Analysis (CDA). Each extract illustrates the pragmatic strategies used by users to mark the boundary between hate speech and free speech. Attention is paid to the illocutionary force of utterances, implicatures, and ideological framing. The extracts have been anonymized and slightly modified to ensure ethical compliance while preserving the original pragmatic intent.

4.1. Extract 1: Reddit User on Immigration"If we can't talk honestly about how illegal immigrants are ruining our neighborhoods, what's the point of free speech?"Speech Act: Assertive with embedded accusation. Gricean Analysis: Violation of the maxim of quantity—"talk honestly" implies previous conversations were dishonest, and "ruining our neighborhoods" is a generalization suggesting negative implicatures. CDA: The speaker frames their speech as a form of truth-telling and silences potential criticism by preemptively aligning themselves with free speech. The discursive strategy constructs immigrants as threats, while positioning the speaker as a silenced truth-bearer, a common rhetorical strategy in right-wing populist discourse (Krzyżanowski, 2020).

4.2. Extract 2: Twitter User on LGBTQ+ Rights"It's just my opinion that pushing this gay agenda in schools is confusing children. Since when did opinions become hate?"Speech Act: Expressive disguised as an assertive. Gricean Analysis: Flouts the maxim of relevance by labeling systemic commentary as mere "opinion," thus masking intent. The phrase "since when did opinions become hate" attempts to deflect responsibility and reframe potential criticism. CDA: The user reframes ideological opposition as personal expression, invoking the right to opinion to delegitimize dissent. This deflective maneuver aligns with broader anti-inclusivity rhetoric that uses free speech as a shield (Wodak, 2015).

4.3. Extract 3: Reddit Response to Accusation"I'm not being racist—I'm just stating facts from government crime stats. Sorry if facts hurt your feelings."Speech Act: Assertive with commissive force.Gricean Analysis: Offends against the maxim of manner in being vague. "Sorry if facts hurt" implies irrationality of sentiment on the part of the critic and makes the speaker appear rational and informed.CDA: "Facts" are rhetorically invoked in a bid to resite discriminatory utterances into the domain of objective facticity. This action is aimed at bolstering structural prejudice in the guise of reason (van Dijk, 1998).

4.4. Extract 4: Twitter Thread on Political Violence"Maybe if these people didn't riot every time they didn't get their way, we wouldn't have to use force. Just saying."Speech Act: Indirect directive with a veiled threat.Gricean Analysis: Conversational implicature from "just saying" is one of neutrality with implication of justification for violence. It goes against the maxim of quality by implying causation without evidence.CDA: The conditional mode ("may be if.") deflects blame to protesters, connoting in justifying state or mob violence while allowing for plausible deniability. This is in accord with authoritarian ideological currents veiled behind informal registers (Fairclough, 1995).

4.5. Extract 5: Reddit Debate on Gender Identity"So now I get banned for saying there are only two genders? That's free speech now?"Speech Act: Directive presented as rhetorical question.Gricean Analysis: Violates maxim of quantity and relation; the figurative status hides the request for confirmation or reinstatement. Ambiguity is used by the user to form a complaint.CDA: The comment repositions platform moderation as censorship, placing the speaker in the role of victim to liberal authoritarianism. Citing being "banned" and the call of "free speech" is part of overarching narratives of anti-PC blowback (Phillips, 2015).

4.6. Extract 6: Twitter Apologia After Accusation"If I offended anyone, that wasn't my intent. I'm just trying to start a conversation."Speech Act: Expressive/apology with implied assertive.Gricean Analysis: The conditional apology ("If I offended") redirects blame and suggests indirectly that the accusation may not have been correct. The "starting a conversation" accusation relocates the speaker as a facilitator and not an aggressor. CDA: This is a typical metapragmatic defense mechanism that transfers the discourse away from the offensive material and to the good faith of the speaker. It supports ideological neutrality, in spite of potential harm coming from the initial speech act (Locher & Watts, 2005).

4.7. Discussion of Patterns Speech Act: Expressive/apology with implied assertive.Gricean Analysis: The conditional apology ("If I offended") blameshifts and suggests indirectly that the accusation may not have been true. The "starting a conversation" accusation reframes the speaker as a facilitator rather than an attacker.CDA: This is a typical metapragmatic defense move that diverts the discourse from the offensive content to the good faith of the speaker. It is compatible with ideological neutrality, regardless of the potential harm radiating from the original speech act (Locher & Watts, 2005).

5. Findings and Discussion This study set out to analyze how online speakers pragmatically navigate and negotiate the oft-blurred boundaries between hate speech and free speech. Employing an eclectic framework combining Speech Act Theory, Gricean Pragmatics, and Critical Discourse Analysis (CDA), discourse analysis of online controversies on Twitter and Reddit revealed a variety of core findings on linguistic and ideological strategies employed by speakers to legitimate, obscure, or resist the intent and impact of inflammatory speech.One central finding is the prevalence of defensive positioning, where users attempt to pre-empt accusations of hate speech by framing their utterances as protected expressions of opinion or fact. Common patterns included disclaimers such as "I'm just stating facts" or "this is my opinion," which, while seemingly neutral, function to cloak ideological stances in objectivity or personal subjectivity. These disclaimers followed on from generalizations about marginalized groups, which essentially redescribed discriminatory statements as valid speech. These strategies follow what van Dijk (1998) has called "semantic moves"—discursive means of protecting speakers from charges of racism or prejudice without abandoning exclusionary ideologies.A further important trend was the use of conversational implicature to deflect responsibility from the speaker onto the addressee. Terms like "sorry if facts hurt your feelings" or "just saying" transgress Grice's maxims of relation and manner so that speakers can deny express purpose but still imply concurrence with provocative or objectionable opinions. This imprecision serves as a guarantee against moderation or ethical responsibility, exploiting platform norms favoring plausible deniability and fragmented context (Marwick & boyd, 2011). This also supports Thomas's (2013) proposal that pragmatic meaning is ordinarily negotiated in what isn't said rather than in what is overtly said.Findings also indicated that the users tend to employ free speech as a metapragmatic frame—a second-order discourse wherein they redescribe their original speech act. For instance, a glaringly offensive or exclusionary statement is not defended on the basis of its content, but on the basis of the speaker's right to make it. This reframing moves the argument from whether a speech act is harmful or not to whether it should be permitted. This sort of framing aligns with Wodak's (2015) observation about how right-wing populist rhetoric employs liberal democratic values (like freedom of speech) to protect ideologically traditional politics. In doing so, free speech is made into a tool—a tool used not simply to defend speech but to stigmatize counter-speech as censorship or hypersensitivity.Another theme emerging is the tactical use of victimhood rhetoric. Some passages include rhetorical questions and passive forms—"So now I get banned for saying there are only two genders?"—that put the speaker in the position of victim of overreach or political correctness.This victimhood rhetoric is the quintessence of populist rhetoric of silenced majorities and adheres to the CDA power negotiation dimension, wherein the speaker tries to upend traditional power structures by reinterpreting dominant cultural demands (e.g., diversity, inclusivity) as oppressive powers (Krzyżanowski2020).Interestingly, the data also showed that users attempted to resolve these conflicts by using speech acts that mitigate conflict, such as "just trying to start a conversation." However, these efforts were often unsuccessful in producing genuine dialogue since the ideologies underlying these original statements

undermined them. Locher and Watts (2005) observe that relational work in online spaces is precarious and easily derailed when power, identity, and moral position are at stake. Together, the findings reveal that the border between free speech and hate speech is not merely legal or moral but discursively demarcated. Users of the internet actively demarcate and redearcate this boundary using pragmatic actions, capitalizing on platform affordances and social ideologies to explain, resist, or counter hate labels. The research testifies that language in online spaces functions not only as a medium of expression but as an ideological battleground, on which variant conceptions of truth, offense, identity, and justice are continuously negotiated.This is a tribute to the requirement for interdisciplinary tools to analyze online discourse, since a single theoretical lens is insufficient to elicit the pragmatic subtlety and the sociopolitical interests involved in employing language in such contested online environments.In sum, the study contributes to a growing body of literature that recognizes the rhetorical density of cyber speech acts and the imperative of understanding how pragmatic vagueness and ideological framing make it possible for hate speech to flourish in the disguise of free expression. Overcoming this challenge will require improved content moderation as much as a more nuanced public debate about the moral limits of free speech in the digital world.

6. Conclusion This study examined pragmatic boundary-marking between hate speech and free speech in online debates through an eclectic framework bringing together Speech Act Theory, Gricean Pragmatics, and Critical Discourse Analysis (CDA). Examining Reddit and Twitter threads on contentious sociopolitical topics, the study identified how online interlocutors use language not only to express opinions but also to construct, defend, or obscure ideologically charged messages. The findings show that online consumers strategically regulate pragmatic norms like implicature, ambiguity, and indirectness to position themselves as freedom defenders or censorship victims, often through speech that stigmatizes or excludes others.Doubtless the most significant observation from the analysis is how free speech is invoked not necessarily as a doctrine but as a rhetorical and ideological tool. When speakers paraphrase controversial or discriminatory messages as free speech ideas or opinions, they can then sidestep criticism and shift the argument from content to perceived rights and identity. This move is usually followed by narratives of victimization and appeals to reason or "facts," which are invoked to delegitimize counterarguments and minimize the harm felt to have been incurred by the initial speech acts.Secondly, the study illustrates ambiguity and plausible deniability as key tactics to navigate platform policy and social censure. This underlines the need to sharpen our insights into digital communication, where pragmatic meaning is in the subtext of the message and not in the letter of the words that are chosen. These findings underscore the shortcomings of strict content moderation policies based solely on overt linguistic signals and demonstrate the merits of contextualized methods rooted in discourse and pragmatic theory.In the end, this book is a contribution towards the broader discussions on digital ethics, public discourse, and the contentious standing of language as a shaper of social norms. It calls for a radical rethinking of the conceptualization and governance of online freedom of expression and calls on scholars, policymakers, and platform designers to get more seriously engaged in the complex dynamics between language, ideology, and power. While internet sites remain central areas of political and cultural discourse, awareness of the pragmatic powers of speech is not only intellectually relevant but socially necessary.

References

· Association of Internet Researchers (AoIR). (2019). Internet research: Ethical guidelines 3.0. https://aoir.org/ethics/

· Barendt, E. (2005). Freedom of speech (2nd ed.). Oxford University Press.

· Blommaert, J. (2005). Discourse: A critical introduction. Cambridge University Press.

· Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage. Cambridge University Press.

· Butler, J. (1997). Excitable speech: A politics of the performative. Routledge.

· Culpeper, J. (2011). Impoliteness: Using language to cause offence. Cambridge University Press.

· Denzin, N. K., & Lincoln, Y. S. (2018). The SAGE handbook of qualitative research (5th ed.). SAGE.

· Fairclough, N. (1995). Critical discourse analysis: The critical study of language. Longman.

· Gelber, K., & Stone, A. (2007). Hate speech and freedom of speech in Australia. Federation Press.

· Graham, S., & Hardaker, C. (2017). (Un)doing gender in political discourse: Women politicians in the UK Parliament. Language & Dialogue, 7(1), 64–86.

· Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Syntax and semantics (Vol. 3, pp. 41–58). Academic Press.

· Heinze, E. (2016). Hate speech and democratic citizenship. Oxford University Press.
· Jhaver, S., Bruckman, A., & Gilbert, E. (2019). Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–27.
· Krzyżanowski, M. (2020). Normalization and legitimization of far-right discourse in mainstream media: The case of Sweden. Discourse & Communication, 14(1), 55–77.
· Locher, M. A., & Watts, R. J. (2005). Politeness theory and relational work. Journal of Politeness Research, 1(1), 9–33.
· Marwick, A., & boyd, d. (2011). The drama! Teen conflict, gossip, and bullying in networked publics. In A. Marwick & d. boyd (Eds.), Youth, identity, and digital media (pp. 1–26). MIT Press.
· Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Information, Communication & Society, 20(6), 930–946.
· Mill, J. S. (2008). On liberty (original work published 1859). Oxford University Press.
· Phillips, W. (2015). This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture. MIT Press.
· Schauer, F. (1982). Free speech: A philosophical enquiry. Cambridge University Press.
· Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Cambridge University Press.
· Thomas, J. (2013). Meaning in interaction: An introduction to pragmatics. Routledge.
· van Dijk, T. A. (1998). Ideology: A multidisciplinary approach. Sage.
· Waldron, J. (2012). The harm in hate speech. Harvard University Press.
· Wodak, R. (2015). The politics of fear: What right-wing populist discourses mean. Sage.
· Wodak, R., & Meyer, M. (2016). Methods of critical discourse studies (3rd ed.). SAGE.