International Journal of Communication 13(2019), 3088–3103 1932–8036/20190005

A Comparative Approach to Social Media Extreme Speech: Online Hate Speech as Media Commentary

MATTI POHJONEN
SOAS University of London, UK

By exploring lessons learned from Ethiopia and Finland, this article challenges two assumptions about online hate speech research. First, it challenges the assumption that the best way to understand controversial concepts such as online hate speech is to determine how closely they represent or mirror some underlying set of facts or state of affairs online or in social media. Second, it challenges the assumption that academic research should be seen as separate from the many controversies that surround online hate speech debates globally. In its place, the article proposes the theory of "commentary" as a comparative research framework aimed at explaining how the messy and complex world of online and social media practices is articulated as hate speech over other ways of imagining this growing problem in global digital media environments.

Keywords: online hate speech, extreme speech, comparative research, commentary, Ethiopia, Finland

Debates around the migration crisis, the resurgence of far-right extremism in Europe and the United States, and online cultures of misogyny, xenophobia, and racism have once again highlighted the growing problem of online hate speech. The stakes could not be higher. Critics have accused hate speech on Facebook for "fueling murderous violence" against minorities in Myanmar and Sri Lanka (Naughton, 2018). Research in Germany has tentatively found that antirefugee rhetoric on social media has led to an increase in hate crimes offline (Müller & Schwarz, 2018). Social media companies and governments are under growing pressure to find new "solutions" to the toxification of online and social media conversations. In the first quarter of 2018 alone, Facebook removed two million pieces of "bad content" from its platform—much of which was done through new mechanisms of algorithmic filtering (Facebook, 2018). This, in turn, has led to civil rights organizations raising concerns about problems of transparency, bias, and accountability involved in the widespread removal of social media content without due process (Article 19 & Privacy International, 2018). How researchers understand online hate speech is arguably one of the most pressing issues facing the future of global digital media research.

But how exactly do we define what online hate speech is? Moreover, what does this contested term imply across a diverse range of countries with often radically different media environments and sociopolitical contexts and histories? Underpinning these questions are a number of theoretical problems that are

Matti Pohjonen: matti.pohjonen@gmail.com
Date submitted: 2018–03–12

notoriously difficult to pin down. How should such expressions of hatred, for instance, be differentiated from the abundance of other types of content produced online? Can this be best achieved through analyzing the surface features of the content produced, which could then be considered indexical of the deeper sentiments, affects, emotions, intentions, motivations, feelings, or attitudes of hatred "based upon the target's being identified with a social or demographic group" (Gagliardone, Gal, Alves, & Martinez, 2015, p. 10)? Or should researchers instead rely on their implicit contextual knowledge and interpretation when determining what constitutes online hate speech (e.g., "I know it when I see it")? Alternatively, how do such expressions of hatred relate to the broader discourse of violence around it? Can the physical or mental harm of online hate speech—discrimination, hostility, or violence—be identified from the content that is shared? Or should the different ways in which audiences interpret the meanings of these expressions of hatred also be factored in when assessing its importance?

A body of scholarly work has begun to disentangle the difficult theoretical questions stirred up by this problem of online hate speech. Critical legal scholars have explored how hateful or otherwise violent speech relates to questions of freedom of expression (Hare & Weinstein, 2011; Herz & Molnar, 2012; Waldron, 2012). Political scientists have, in turn, analyzed the use of hateful language by extremist political movements (Brindle, 2016). Peace and conflict studies scholars have linked online hate speech debates to the dynamics of violent conflict (Buyse, 2014). Computer scientists have explored computational methods for identifying expressions of hatred from the abundance of "big data" found online and in the social media (see Burnap & Williams, 2015; Davidson, Warmsley, Macy, & Weber, 2017).

In this article, however, I depart from these approaches to foreground a relatively understudied perspective to online hate speech debates. Rather than trying to define what constitutes online hate speech, or even what its cultural, political, and social consequences are, I ask instead, What does it mean to define something as online hate speech in the first place? I do this by exploring theoretically the different ways researchers negotiate—through their theoretical frameworks, methodological choices, and everyday research practices—the contested definitions and political controversies that are superimposed on this object of study. By doing this, the article challenges two assumptions about online hate speech research in a comparative global context.

First, it challenges the assumption that the concepts researchers use, academic or otherwise, should be primarily seen as representing or mirroring some underlying reality, set of facts, or state of affairs. In its place, I argue that the concepts that researchers use must also function performatively; that is, they are used to enact different outcomes in the situations they are deployed in. From this perspective, how researchers define what online hate speech is cannot be limited only to what this concept ostensibly means, or its correspondence to some set of underlying practices found online or in the social media, but has to do as much with what the concept tries to achieve across a diverse range of situations. Second, it challenges the assumption that academic practice, especially when involving controversial concepts such as online hate speech, cannot be seen as separate from the social and political antagonisms that surround it. In its place, the article therefore advances a "bi-focal" or "doubly critical" (Pohjonen, 2014) research approach to the contradictory meanings that online hate speech is given in different global contexts. An integral part of this includes developing a critical ethnographic sensibility to the different practices of how the discourse of online hate speech itself is strategically deployed across a variety of political debates and, as important, how

researchers themselves are inextricably situated as participants in the sociopolitical assemblages that constitute these debates (Delanda, 2016).

The argument proceeds in four parts. The first part examines the theoretical challenges involved in defining online hate speech from a comparative global perspective. The second part outlines an alternative theoretical framework for online hate speech research based on approaching it as a form of media commentary through which the messy world of everyday online and social media practices is given retroactive meaning and closure. The third part illustrates the argument by comparing two research projects on online hate speech in two distinctly different sociopolitical contexts and media environments: Ethiopia and Finland. The article concludes by making the case for the concept of "extreme speech" as an anthropological qualifier to online hate speech debates—a concept that was developed to allow more theoretical flexibility to research the specific cultural contexts and situated practices surrounding this phenomenon globally.

How to Do Things With Online Hate Speech?

The difficulty in defining what online hate speech is can be partially explained by the multiple overlapping discourses that operate on the concept. Brown (2017a, 2017b) argues that a critical distinction needs to be made between the use of hate speech as a legal concept and its use as an ordinary concept. By this, he means that the critical legal-philosophical debates that inspired the earlier important scholarly interventions into hate speech debates have now also spilled over to their "ordinary" uses. The exact theoretical terminology developed to explore the tenuous relationship between hateful speech and freedom of expression has been thus unwedded from its original purpose. Its use in public and political debates has less to do with legal elucidation or philosophical clarity and more to do with rhetorically advancing ideological viewpoints. Brown (2017a) writes,

> the term "hate speech" has been perhaps most often associated with liberal progressives, or people on the left of politics—who use it to highlight and problematize speech that they view as racist, xenophobic, homophobic, Islamophobic, misogynistic, disablist, or in some other way targeted at minority groups in ways that supposedly violate ideals of respect, solidarity, tolerance, and so forth. By contrast, many political and religious conservatives repudiate such uses of the term, and view them simply as crude attempts to close down meaningful debate on what they believe are the evils of open-border policies, the failures of multiculturalism as a social experiment, the lamentable decline of traditional moral values, political correctness gone mad, and so on. (p. 425)

Global digital media researchers working with online hate speech therefore must negotiate its uses both in legal doctrine and criminal law, and in its more promiscuous appropriation for political rhetoric in public debates. This conceptual ambiguity has precipitated a number of satellite concepts, which have each tried to differentiate more narrowly defined subsets of online expressions of hatred. Benesch (2014), for instance, uses the concept of *dangerous speech* to highlight the risks of real-world violence and genocide associated with hateful speech acts in situations of violent conflict. Buyse (2014) talks about *fear speech* when focusing on the sociopsychological dynamics of fear-mongering associated with such expressions of hatred. Saleem, Dillon,

Benesch, and Ruths (2017) propose *hateful speech* as a conceptual shortcut for aiding computational detection of expressions of hatred from the abundance of other digital traces generated in online and social media conversations. Other terms, such as *anti-social media*, *offensive speech*, *excitable speech*, *online vitriol*, *cyber-bullying*, *micro-aggression*, or *inflammatory language*, have also been widely used as metonymical substitutes for the concept of online hate speech as an effort to buffer the more normative legal conjunctions associated with it (Gagliardone, 2019).

The concepts that researchers use are important because they influence the outcome of the research (e.g., what conceptual criteria the research framework needs to use to differentiate relevant expressions of hatred from all the other social media conversations "out there") and how others engage with the research (e.g., despite its conceptual ambiguities, the term *hate speech* still carries more rhetorical value in public and political debates than other terms such as *inflammatory language* or *offensive speech*). Moreover, research has also shown that when working with contested concepts such as online hate speech, it is often difficult for researchers to reach intercoder agreement on how exactly to identify what types of speech acts should fall under this category. Kwok and Wang (2013) and Ross et al. (2016) have demonstrated how, even with the use of human annotators doing the categorization, what becomes classified as hate speech significantly depends on the backgrounds of the researchers doing the annotation, such as gender, ethnicity class, political orientation, and age. This process is also often complicated by the existence of coded linguistic forms such as jokes, innuendo, irony, metaphors, or double meanings associated with subcultural communities online. Identifying the subtle cultural nuances in the use of everyday language, therefore, requires significant levels of familiarity with the online cultures behind such expressions of hatred, which makes the generalization of speech categories across different cultural contexts problematic (Udupa, 2017).

The difficulties involved in defining what online hate speech is, however, also highlight a more fundamental philosophical question than simply choosing the right conceptual framework or methodology for research (e.g., even using long-term ethnographic observation and sensitivity to the emic insights of online cultures will not solve all the challenges in defining online hate speech). This question has to do with whether controversial concepts such as online hate speech can ever have one determinable meaning that researchers should strive for, or should at least try to approximate, in their research frameworks. Contrary to this orthodoxy, for instance, Brown (2017b) argues that concepts such as hate speech should be instead approached through Wittgenstein's theory of "family resemblance" concepts—that is, concepts that do not, or cannot, possess a singular or universal meaning outside their uses in different situations. He writes,

> What I am claiming, in other words, is not simply that a variety of different things can count as hate speech . . . I am claiming that the term "hate speech" has more than one meaning. Of course, it has become something of a cliché to assert that there is profound disagreement about the meaning of the term "hate speech," disagreement not only among legal scholars and legal professionals but also among ordinary language users. But my claim is not that people disagree about what the correct definition of the term is; after all, that would be consistent with one of the definitions being correct and there actually being a single meaning. Instead, what I am claiming is that the term "hate speech" is systematically ambiguous; which is to say, it carries a multiplicity of different meanings. (p. 564)

This kind of antiessentialist view of language runs counter to the more commonsensical view of global digital media research according to which the utility of concepts is determined based on how accurately they are able to represent some underlying reality or state of affairs (e.g., the view that the definition of online hate speech should be understood based on how it reflects a set of activities found online). In this alternative view, however, the meaning of concepts needs to be also understood in their performative context. That is, rather than only being used in an effort to describe the messy world "somewhere out there," they are also used strategically to enact specific outcomes and interventions in the situations in which they are deployed. Such a more pragmatic view of language use has many theoretical precedents from Wittgenstein's "language games" (Wittgenstein, 2009), Austin's (1962) "speech-act theory" (see also Butler, 1997, Hartley's (1995) "intervention analysis," the concept of "articulation" in the cultural studies (Slack, 1996), the concept of "actants" in actor-network theory (Latour, 2007), or even the theory of "order-words" in the works of Deleuze (1987).

This complex debate around the philosophy of language is, of course, beyond the scope of this article (Deleuze, 1994; Rorty, 2017). Yet what this alternative approach to online hate speech debates implies is a shift in analytical register away from seeing online hate speech as a "transparent concept" that we can "glance through" or "a simple window to reality" (Boromisza-Habashi, 2013, p. 2), and toward a more nuanced appreciation of the situated everyday practices through which concepts are given their multiplicity of meanings. Unavoidably, these also include the practices of the researchers who work with online hate speech in their research.

What I suggest in this article, therefore, is that if there are indeed no essential meanings behind the concept of online hate speech, one way that research can negotiate these differential meanings that online hate speech acquires in its different uses around the world is to develop a kind of doubly critical or bi-focal perspective toward these "different registers of truth in their articulation with each other" (Morley, 2006, p. 32). As critical anthropologists have argued, the concepts that researchers use to explain the world should be seen also as "social facts," facts that are inextricably linked to the sociocultural contexts of their production and the various shifting historical power relationships that have influenced how they circulate across a range of political, legal, and academic contexts (Rabinow, 1986). Instead of a transparent "mirror to reality," what we find is a complex series of "cultural translations" through which researchers negotiate the conflictual and sometimes contradictory truth claims and worldviews of the participants involved in the research (Asad, 1986; Hobart, 1996).

From this perspective, the analytical focus of online hate speech research shifts to a different ontological register: It becomes about the myriad practices through which the messy world of online practices and left-behind digital traces are represented as online hate speech over all other possible ways of imagining this growing problem of our global digitally mediated communication.

Online Hate Speech as a Form of Media "Commentary"?

One way to approach the ongoing debates on online hate speech is to see them as forms of media commentary. Hobart (2001) argues that media studies has historically faced the difficulty of theoretically pinning down its object of study. This is because the research has presupposed that there is something

substantial that underlies the media texts, production practices, or audience interpretations that the researcher is privy to interpret—whether this something is the "meaning" of the media text, the "culture" influencing the practices, the "ideology" of the audiences interpreting the media texts, or some other category that would preexist its mediation in academic, public, or political debates. Against these essentialist approaches, however, Hobart suggests that it makes more sense to approach media text and practices as underdetermined. This is to say, the multiplicity of sociopolitical assemblages that constitute global media today have always more meanings superimposed on them than any research account is able to capture. Moreover, because of this fundamental impossibility of providing closure to what in reality consists of an almost infinitely complex set of social, political, and cultural practices and histories, one key part of contemporary media practices and representations involves practices whose primary purpose is to comment on what the significance of these prior media practices is in the first place and how people should understand them.

This concept of media "commentary" builds on Foucault's analysis of the discursive mechanisms through which the production of knowledge is regulated in contemporary societies. Foucault (1981) wrote that "the commentary's only role, whatever the technique used, is to say at last what was silently articulated 'beyond,' in the text" (p. 57). Such commentaries, therefore, consist of the different discursive practices through which earlier texts (e.g., such as expressions of hatred found online on any given day) are given retroactive meaning despite there being no such singular meanings behind them to begin with (e.g., articulating the diverse range of hateful expressions of hatred online *as* hate speech over other ways of imagining the problem). Foucault argued that such

> commentary exorcises the chance element of discourse by giving it its due: it allows us to say something other than the text itself, but on condition that it is this text itself which is said, and in a sense completed. The open multiplicity, the element of chance, are transferred, by the principle of commentary, from what might risk being said, on the number, the form, the mask, and the circumstance of the repetition. The new thing lies not in what is said but in the event of its return. (p. 58)

Somewhat paradoxically, then, the purpose of such media commentaries is not found only in what these commentaries say or mean, but also in their attempt to provide retroactive closure to the overflow of meaning that characterizes the messy world of social and cultural practices through the endless of repetition of commentaries about what the "real" meaning of the phenomena under consideration is or how it should be understood.

In other words, then, from the perspective of this kind of antiessentialist approach to online hate speech debates, such commentaries consist of those moments in which the meaning of online hate speech itself is debated and contested in different situations globally and thus given its significance. Moreover, what is relevant about such media commentaries is that many of them take place in the digital media itself. Approaching online hate speech as a form of media commentary, therefore, can open up a new empirical object of study for researchers interested in empirically researching debates on online hate speech. That is, instead of starting the research from the Sisyphean task of trying to provide a definition for online hate speech, research can instead step back from the politicized debates and controversy—

even if strategically and temporarily—and foreground a more critical sensibility toward the everyday practices through which such political meanings are produced and contested in global digital media commentaries.

My interest in this kind of critical meta-methodological perspective to the global discourse of online hate speech, and the academic research thereof, derives from research projects in which I have been involved in Ethiopia and Finland. The next section specifies the theoretical argument by illustrating its key points through comparing online hate speech research across these two distinctly different sociopolitical contexts and media environments.

Researching Online Hate Speech in Ethiopia

Despite rhetoric about its importance as a forum for political participation, Internet accessibility levels in Ethiopia remain some of the lowest in the world. Ethiopia has maintained (at least until recently) one of the most extensive systems of online censorship and surveillance in Africa, and its government has routinely arrested journalists and bloggers for expressing critical voices online. Within this context, the discourse of online hate speech in Ethiopia has been linked to the political antagonisms in the country and invoked by both the government and the opposition to express a panoply of other grievances (Gagliardone, 2017; Legesse, 2012). Two examples illustrate how contested the meaning of online hate speech was in Ethiopia during the time of our research.

When we posted an advertisement online looking for research assistants, Tigray Online—an online news site considered by many to be associated with the government—published an article titled "Oxford University Wants to Study Ethiopian Election Time Propaganda and Online Debates, but Why?" The writer of the article expressed concerns about what the political motives of online hate speech research in the Ethiopia context were:

> I cannot help but feel why spend all this money to identify and understand Ethiopian hate speech online. . . . This move *can enable foreign forces, especially those who have been working extra hours to incite color revolution in Ethiopia, join the online hate speeches we are not so proud of ourselves and guide them into directions we as Ethiopians would not like them to go.* No matter how far apart and antagonistic our political stands are, we all know better than inviting foreigners to come and mess things up even further. (Gebru, 2014, paras. 5–6, emphasis added)

The online article further commented on how this research should be positioned into the longer history of "foreign forces" working to destabilize the government by unfairly highlighting the political tensions in the country. The author of the article also cited an earlier report we had published, which had identified some of the challenges involved in researching online hate speech in such politically polarized countries. In the report, we had suggested that one of the aims of online hate speech research should be to provide a "neutral platform for mediation through which the antagonisms underpinning hate speech could be better identified and steps be taken to mitigate them" (Gagliardone, Pohjonen, & Patel, 2014, p. 37). The author of the Tigray Online article, however, argued that, even with the best intentions of providing such a neutral

platform, the study would nonetheless have negative effects in Ethiopia because it would get "used by forces that would like to incite violence in the country" (Gebru, 2014, para. 11).

Conversely, following an interview about our research results in a diasporic Ethiopian online news site, an article titled "Ethiopia: Is University of Oxford Cooking a Study?" was published in ECAFD Online, a news forum associated with the Ethiopian and diasporic opposition. The article accused the research of selling out to the government:

> From all countries to pick Apartheid Ethiopia with the lowest internet penetration controlled by TPLF[1] intelligence agency that *as a policy promote hate and violence among Ethiopians is a tragedy by its own.* (Debalke, 2016, para. 17, emphasis added)

The article, which was shared 351 times, further argued that the primary purpose of the research was to serve the interests of the Ethiopian regime.

These two examples illustrate how overdetermined the meaning of online hate speech was in Ethiopia during the time of our research. From the government's side, research on online hate speech was linked to the history with foreign forces to destabilize the country. On the opposition side, researchers were accused of working for the same government. Publicly summoning the term *online hate speech* thus brought into ambit all kinds of diametrically opposed commentaries about its significance that had little to do with the intentions or motivations of our original research. Regardless of what the research framework was, or even what its results showed, the research was clawed back into the preexisting political grievances in Ethiopia and the broader history of political struggle and conflict in the country.

It was within this context that we had to negotiate the many contested meanings given to online hate speech in Ethiopia while conducting our research. One way to do this was to host workshops where the significance of online hate speech itself could be debated among participants from different political orientations—each with radically different ideas of what online hate speech meant in the Ethiopian context, what should be done about it, and, especially, who was to blame for it. What started out as an exploratory research project into an understudied topic quickly became associated with a host of other meanings—meanings that had little to do with how this object of study is commonly understood in Western debates. By strategically leaving the definition of online hate speech as open as possible, online hate speech was thus commented on and understood by the workshop participants to be as much a manifestation of the underlying social and political conflict in Ethiopia as it was about determining what types of speech acts should be excluded from the legitimate space of political expression and who was to blame for it.

In our follow-up project in Ethiopia, we tried to incorporate these antagonistic meanings into its research framework. This was done through a number of deliberate theoretical and methodological interventions. First, we developed a sampling strategy, trying to contextualize the politically charged debates on online hate speech by situating them within the broader communicative milieu and cultures of communication of Ethiopian and diasporic online spaces that went beyond a simplistic binary understanding

[1] TPLF refers to the Tigrinya People's Liberation Front, one of the dominant political groups in Ethiopia.

of hate speech/not hate speech. Rather than focusing only on statements categorized as online hate speech, the research resulted in the creation of a sample frame composed of more than 1,000 Facebook pages that reflected a diversity of popular Ethiopian and diasporic online conversations. Second, the conceptual framework used in the research purposefully moved away from the legal-normative approaches to hate speech to foreground the different types of communicative relationships that form in online conversations. This was done through conceptually categorizing statements based on whether they facilitated (going toward) or hindered (going against) dialogue and engagement among the interlocutors involved. Statements that went against thus included statements with "conflict-producing" or "conflict maintaining behavior," such as "attacking another speaker or a specific group by belittling, challenging, provoking, teasing them maliciously, or explicitly threatening them" (Gagliardone et al., 2016, p. 17). Statements that went toward, on the contrary, included statements that tried to build communicative relationships through "acknowledging another person or group's position, offering additional information about the topic being discussed, joking (in a non-hostile teasing way), and creating engagement and conversation with the other members in the discussion" (p. 17).

This kind of pragmatic approach to the definition of online hate speech was aimed as much toward promoting cross-political dialogue as it was about taking sides or blaming any of the parties involved. It also allowed the research to produce new empirical insights into the nature of online conversations in Ethiopia. Using this research framework, we analyzed more than 13,000 messages from this sample frame over a four-month period (see Gagliardone et al., 2016, pp. 12–22). What was surprising about approaching debates on online hate speech from this perspective was that social media conversations in Ethiopia, in addition to containing traces of the ethnic and political conflict in the country, also seemed to also promote spaces for constructive political engagement. Although the worst kinds of hate speech or dangerous speech were still found, they were uttered by people who were anonymous or had little influence. More crucially, the research also found no hateful speech acts in which the speakers had the actual means to carry out the threat of violence in real-world situations. While the findings of the research were, of course, particular to the research questions posed, the methodology used, and the idiosyncratic political situation in Ethiopia at the time of research, approaching online hate speech from such a holistic communicational perspective took away some of the rhetorical power of the government's argument for further censoring social media conversations in Ethiopia as a means to prevent its imagined dangers.

Researching Online Hate Speech in Finland

What is curious when approaching online hate speech from such a comparative perspective (e.g., how the lessons learned in one context can be applied to a different context) is how differently debates are discursively framed depending on the part of the world that is the focus of the research (Jackson, 2012). With the risk of caricaturing or simplifying what in reality consists of a range of different and often incommensurable theoretical positions, when the question is about online hate speech in the West, the problem seems to be framed more in terms of a "discourse of pathology" that divides the social world into a mainstream center and an extremist periphery. From this perspective, the problem of online hate speech becomes more about identifying where the boundary between acceptable and unacceptable speech is drawn and who should be excluded from the legitimate sphere of political expression. However, when the debates focus outside the liberal West, and in countries such as Ethiopia, the problem of online hate speech is often

framed more from the perspective of ethnic or political conflict and its mediation. The underlying metaphor here, in contrast, divides the social world into two (or more) parallel sides involved in a conflict situation and a shifting zone of engagement through which peaceful solutions to the conflict could be found, and perhaps pockets of "spoilers" who are left out of the peace process because they are not willing to enter a peaceful solution or actively try to provoke violence to derail solutions to conflict (Stedman, 1997).

The problem of online hate speech, therefore, becomes as much about finding ways to mitigate the social and political tensions underlying such expressions of hatred as it is about determining what kinds of speech acts are acceptable and what is not—as perhaps has been more common in the historical genealogy of hate speech debates in the West revolving around questions of freedom of speech.

In my second research project, I thus aimed to explore how these lessons learned in Ethiopia could apply to the European context during the so-called 2015–2016 migrant/refugee crisis, and to Finland in particular. What was surprising about approaching social media conversation this way was how prolific hateful and violent expressions against immigrants had become during the migrant/refugee crisis. All the characteristics of the worst kinds of hate speech we had previously identified in situations of violent ethnic and political conflict—comparing people to animals or vermin, explicit calls to violence and accusations in a mirror—could be found in abundance in Finnish social media and online spaces.

This raised an interesting theoretical dilemma that I wanted to explore in my research. Indeed, if such expressions of hatred in what has often been considered the "safest country in the world" had become more abundant, aggressive, and violent than what our research had found in Ethiopia, a country with a long history of ethnic and political conflict and violence, how should we best theoretically approach this nebulous relationship between online speech and the discourse of violence surrounding it? Said differently, how should we understand online hate speech in a situation in which the tenuous relationship between online speech and the discourse of violence seemed to follow a kind of counterintuitive logic? That is, there were more expressions of hatred online in a country that was, at least on the surface, more peaceful, at least in terms of fewer outbreaks of violence. What kind of comparative research frameworks and operational definitions of hate speech would, then, take into count this somewhat counterintuitive logic?

During the so-called migrant/refugee crisis in 2015–2016, debates on hate speech in Finland had become polarized between the anti-immigrant and the antiracist groups, each holding diametrically opposed viewpoints on the question of the arriving refugees. On the one side of the political spectrum was the popular *Rajat Kiinni* (Close the Borders) Facebook group, which had become a notorious for its antirefugee sentiments and hateful tone of conversation. On the opposing side, the *Rasmus* group (Finland´s national network and association working against racism and xenophobia and promoting equity and human rights) had adopted an explicitly antiracist position. A big part of the interaction among these opposed groups consisted of promoting screenshots accusing each other of promoting hate speech or defending against these accusations.

Given this context of social media polarization in Finland, again, instead of trying to define what was meant by these ongoing accusations of hate speech in these debates a priori, in my research, I decided to instead explore how the concepts related to debates on hate speech themselves were given often

conflictual meaning by the participants involved in the debates. What was interesting about approaching the debates on online hate speech in Finland this way was how antagonistically the key terms associated with the migrant/refugee crisis were commented on in these conversations. For instance, words closely associated with the term *pakolainen* (refugee) in the antirefugee/immigration Facebook group Rajat Kiinni included a host of negative connotations, such as "parasite" or "welfare refugee." On the contrary, in the antiracist group Rasmus, this term was more closely associated with words connoting forced movement and the need to help.

Crucially, when the research explored what words were associated with *vihapuhe* (hate speech) itself in the two opposed groups, this term was defined and also understood in radically different ways. In the antiracist group Rasmus*, vihapuhe* was closely associated with terms such as *violence*, *xenophobia*, *discrimination*, and *zero tolerance*. On the contrary, in the anti-immigration/refugee Facebook group Rajat Kiinni*, vihapuhe* was closely associated with terms connoting the expression of opinions and of being accused and judged. These findings were based on computational text mining of word associations across close to 500,000 comments found in these two groups and confirmed by a longer term digital ethnographic observation of the types of posts and comments found in these different groups. When members of the antiracist Rasmus group referred to debates on online hate speech, usually they did so in line with a more mainstream criticism of racist speech, fascism, and the rise of the anti-immigration far right. Conversely, when members of the antimigrant/refugee group Rajat Kiinni referred to debates on online hate speech, this was framed more as an attempt by the mainstream or the liberals to censor their opinions or hide the "real truth" about immigration (Pohjonen, 2018).

One conclusion that I was able to draw from this comparative experiment was that even the key concepts, including the definition of hate speech itself, are understood in radically different ways by the participants involved. While in no way condoning the vitriolic expressions of hatred abundantly found in anti-immigration groups in Finland or the emotional harm they can cause, I concluded that these definitions need to be incorporated into the research framework to attain a better understanding of the social and political antagonisms that generate such expressions of hate in the first place, and even what the grievances or the *jouissance* and "fun" driving these conversations are (Udupa, 2019). In taking such a bi-focal or doubly critical approach to online hate speech debates, one must, therefore, remain both critical of what is being expressed in these vitriolic debates and acknowledge that there are often radically different truth-claims by the participants involved; this must always be negotiated while conducting research, even in relatively homogenous societies like in Finland.

Discussion

What, then, can these two different research experiments tell us more broadly about such comparative approaches to online hate speech? As the examples from Ethiopia and Finland suggest, underlying online hate speech debates is a complex panoply of different sociopolitical dynamics, conflictual dynamics, and histories that risk being obviated by the more common legal-normative understandings of online hate speech. In both examples, through leaving the definition of online hate speech as open-ended as possible, in my research I instead chose to use it as a kind of "empty signifier" through which alternative viewpoints could be raised in these emotionally charged and overdetermined political debates (Laclau &

Mouffe, 1985). In Ethiopia, the aim of this was more to foreground questions of political participation and engagement evoked by debates on hate speech. In the case of Finland, the aim was to move the research framework beyond a common legal-normative definition of online hate speech to instead explore the sociopolitical tensions, conflictual dynamics, and contested meanings that underpinned such eruptions of hatred online during the migrant/refugee crisis in a relatively peaceful country with fewer instances of outbreaks of mass offline violence.

Approaching online hate speech debates from such a bi-focal or doubly critical perspective does not, of course, mean that all media commentaries or definitions of online hate speech should be given moral equivalence, nor should the viewpoints of the targets of this speech or the harm caused by it be dismissed from the accounts. On the contrary, as is the case with any other political debate, there are always complex questions of power and practices of exclusion involved in how such media commentaries around online hate speech are produced and who benefits and suffers from them. In the case of Ethiopia, these power imbalances were linked to the post-civil war history of the country and the grievances raised against the ruling government. In the case of Finland, there are also unequal power relations involved in how online hate speech is articulated by immigrants, organized racist groups, academic researchers, the police, and legislative bodies. Rather, what this comparative research framework suggests is that these commentaries, and their rightful criticism, must always be situated within these sociopolitical assemblages where they are inextricably embedded to both better understand what the stakes are and find better ways to counter them.

In conclusion, then, what I have argued in this article is that such a comparative approach to online hate speech debates has two potential implications for online hate speech research more broadly. First, approaching online hate speech debates as a form of media commentary potentially allows the research to empirically focus on the situations in which these differential meanings associated with this concept of online hate speech are produced, contested, and re-produced globally today and how these meanings are culturally translated and transmuted across different contexts and situations. The question here thus becomes less about what online hate speech is and more about why and where something is represented as online hate speech in the first place over other possible ways of imagining this growing problem of contemporary global digital media environments. The second implication derives from the first. It has to with the question of why researchers choose the specific concepts they use in their research in the first place, especially when dealing with contested topics such as online hate speech that arguably have no singular way to be represented. That is to say, if there are no essential meanings behind concepts to begin with, how researchers use these concepts also must be understood in their performative contexts, through the strategic goals that the use of these concepts is enacted to achieve in research situations and contexts that are unavoidably linked to the complex political contexts around them.

The concept of extreme speech advanced in this Special Issue was developed as an anthropological qualifier to global debates on online hate speech, with the strategic goal of allowing researchers to better take into account the situated practices and cultural contexts behind the many different kinds of hateful online speech cultures that exist globally and the people behind them. Approaching online hate speech debates as form of media commentary thus builds on this research agenda to also foreground and highlight the different ways in which the multiplicity of online hate speech debates, and the many controversies around them, are given their significance globally and for what kinds of purposes.

References

Article 19 & Privacy International. (2018). *Privacy and freedom of expression in the age of artificial intelligence*. Retrieved from https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf

Asad, T. (1986). The concept of cultural translation in British social anthropology. In J. Clifford & G. E. Marcus (Eds.), *Writing culture: The poetics and politics of ethnography* (pp. 141–165). Los Angeles, CA: University of California Press.

Austin, J. L. (1962). *How to do things with words: The William James lectures delivered at Harvard University in 1955*. Oxford, UK: Clarendon Press.

Benesch, S. (2014). *Countering dangerous speech: New ideas for genocide prevention* (Working paper). Washington, DC: United States Holocaust Memorial Museum. Retrieved from https://dangerousspeech.org/countering-dangerous-speech-new-ideas-for-genocide-prevention/

Boromisza-Habashi, D. (2013). *Speaking hatefully: Culture, communication, and political action in Hungary*. University Park, PA: Penn State University Press.

Brindle, A. (2016). *The language of hate: A corpus linguistic analysis of white supremacist language*. New York, NY: Routledge.

Brown, A. (2017a). What is hate speech? Part 1: The myth of hate. *Law and Philosophy*, *36*, 419–468.

Brown, A. (2017b). What is hate speech? Part 2: Family resemblances. *Law and Philosophy*, *36*, 561–613.

Burnap, P., & Williams, M. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modelling for policy and decision making. *P&I Policy and Internet*, *7*(2), 223–242.

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

Buyse, A. (2014). Words of violence: "Fear speech," or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, *36*(4), 779–797.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). *Automated hate speech detection and the problem of offensive language*. Retrieved from https://arxiv.org/pdf/1703.04009.pdf

Debalke, T. (2016, July 11). Ethiopia: Is University of Oxford cooking a study. *ECADF Ethiopian News*. Retrieved from https://ecadforum.com/2016/06/11/is-university-of-oxford-cooking-a-study-titled-mechachal

Delanda, M. (2016). *Assemblage theory.* Edinburgh, UK: Edinburgh University Press.

Deleuze, G. (1987). *A thousand plateaus: Capitalism and schizophrenia*. Minneapolis, MN: University of Minnesota Press.

Deleuze, G. (1994). *Difference and repetition.* New York, NY: Columbia University Press.

Facebook. (2018, May 2). *F8 2018: Using technology to remove the bad stuff before it's even reported*. Retrieved from https://newsroom.fb.com/news/2018/05/removing-content-using-ai/

Foucault, M. (1981). Order of discourse. In R. Young (Ed.), *Untying the text: A post- structuralist reader* (pp. 48–79). London, UK: Routledge & Kegan and Paul.

Gagliardone, I. (2017). *The politics of technology in Africa: Communication, development, and nation-Building in Ethiopia*. Cambridge, UK: Cambridge University Press.

Gagliardone, I. (2019). Defining online speech and its "public lives": What is the place for "extreme speech"? *International Journal of Communication, 13,* this Special Section.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. UNESCO Series on Internet Freedoms. Paris, France: UNESCO Publishing.

Gagliardone, I., Pohjonen, M., & Patel, A. (2014). *Mapping and analysing hate speech online: Opportunities and challenges for Ethiopia*. The Programme in Comparative Media Law and Policy, University of Oxford, and Addis Ababa University. Retrieved from http://pcmlp.socleg.ox.ac.uk/wp-content/uploads/2014/12/Ethiopia-hate-speech.pdf

Gagliardone, I., Pohjonen, M., Zerai, I., Beyene, Z., Aynekulu, G., Stremlau, N., . . . Gebrewolde, T. M. (2016). *MECHACHAL: Online debates and elections in Ethiopia. From hate speech to engagement in social media*. The Programme in Comparative Media Law and Policy, University of Oxford, and Addis Ababa University. Retrieved from https://papers.ssrn.com/sol3/papers.cfm ?abstract_id=2831369

Gebru, B. (2014, November 14). Oxford University wants to study Ethiopians election time propaganda and online debates, but why? *Tigray Online.* Retrieved from http://www.tigraionline.com/articles/oxford-study-election.html

Hare, I., & Weinstein, J. (2011). *Extreme speech and democracy*. Oxford, UK: Oxford University Press.

Hartley, J. (1995). *Tele-ology*. London, UK: Routledge.

Herz, M., & Molnar, P. (2012). *The content and context of hate speech: Rethinking regulation and responses*. Cambridge, UK: Cambridge University Press.

Hobart, M. (1996). Ethnography as a practice: Or the unimportance of penguins. *Europeae, 2*(1), 3–36. Retrieved from http://eprints.soas.ac.uk/7084/1/Ethnography_as_a_practice_-_published_version.pdf

Hobart, M. (2001, April). *Loose cannons: Commentary as the missing object in Indonesian media studies*. Paper presented at the VA/AVMI Symposium on Media Cultures in Indonesia, Leiden University, Leiden, The Netherlands. Retrieved from https://www.academia.edu/35623269/Loose_canons_commentary_as_the_missing_object_in_Indonesian_media_studiese

Jackson, R. (2012). Unknown knowns: The subjugated knowledge of terrorism studies. *Critical Studies on Terrorism*, *5*(1), 11–29.

Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against Blacks. *AAAI'13: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 1621–1622. Retrieved from https://pdfs.semanticscholar.org/db55/11e90b2f4d650067ebf934294617eff81eca.pdf

Laclau, E., & Mouffe, C. (1985). *Hegemony & socialist strategy*. London, UK: Routledge.

Latour, B. (2007). *Reassembling the social: An introduction to actor-network-theory*. Oxford, UK: Oxford University Press.

Legesse, Y. M. (2012). Shielding marginalized groups from verbal assaults without abusing hate speech laws. In M. Herz & P. Molnar (Eds.), *The content and context of hate speech* (pp. 352–377). Cambridge, UK: Cambridge University Press.

Morley, D. (2006). Globalisation and cultural imperialism revisited: Old questions in new guises. In D. Morley & J. Curran (Eds.), *Media and cultural theory* (pp. 30–43). London, UK: Routledge.

Müller, K., & Schwarz, C. (2018). Flaming the flames of hate: Social media and hate crime. *SSRN*. Retrieved from https://ssrn.com/abstract=3082972

Naughton, J. (2018, April 29). Facebook's global monopoly poses a deadly threat in developing nations. *The Guardian.* Retrieved from https://www.theguardian.com/commentisfree/2018/apr/29/facebook-global-monopoly-deadly-problem-myanmar-sri-lanka

Pohjonen, M. (2014). *In media res: The problem of cultural translation of international news in Mumbai, India* (Unpublished PhD thesis). School of Oriental and African Studies, University of London, London, UK. Retrieved from https://eprints.soas.ac.uk/20351/

Pohjonen, M. (2018). *Horizons of hate: Comparative approach to online hate speech*. Retrieved from https://www.voxpol.eu/download/vox-pol_publication/Horizons-of-Hate.pdf

Rabinow, P. (1986). Representations are social facts: Modernity and post-modernity in anthropology. In J. Clifford & G. E. Marcus (Eds.), *Writing culture: the poetics and politics of ethnography* (pp. 235–261). Los Angeles, CA: University of California Press.

Rorty, R. (2017). *Philosophy and the mirror of nature* (30th anniversary ed.) Princeton, NJ: Princeton University Press.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *Proceedings of NLP4CMC III. Bochumer Linguistische Arbeitsberichte*. Retrieved from https://arxiv.org/pdf/1701.08118.pdf

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. In *Proceedings of First Workshop on Text Analytics for Cybersecurity and Online Safety.* Retrieved from https://arxiv.org/abs/1709.10159

Slack, J. (1996). Theory and method of articulation in cultural studies. In D. Morley (Ed.), *Stuart Hall: Critical dialogues in cultural studies* (pp. 112–131). London, UK: Routledge.

Stedman, J. (1997). Spoiler problem in peace processes. *International Security*, *22*(2), 5–53.

Udupa, S. (2017). Gaali cultures: The politics of abusive exchange on social media. *New Media and Society, 20*(4), 1506–1522.

Udupa, S. (2019). Nationalism in the digital age: Fun as a meta-practice of extreme speech. *International Journal of Communication*, *13*, this Special Section.

Waldron, J. (2012). *The harm in hate speech*. Cambridge, MA: Harvard University Press.

Wittgenstein, L. (2009). *Philosophical investigations* (4th ed.). London, UK: Wiley-Blackwell.