

First-person verbal aggression in YouTube comments

Author accepted manuscript

Ylva Biri, University of Helsinki

Laura Hekanaho, University of Helsinki

Minna Palander-Collin, University of Helsinki

Version of record:

Biri, Ylva; Hekanaho, Laura & Palander-Collin, Minna (2023). First-person verbal aggression in YouTube comments on sexism. In Ermida, Isabel (ed.). *NetLang: The Language of Cyberbullying: Forms and Mechanisms of Online Prejudice and Discrimination*, pp. 107-137. Palgrave Macmillan.
<https://link.springer.com/book/10.1007/978-3-031-38248-2>

Abstract. To contribute to literature on hateful speech in social media communication, this paper analyses verbal aggression in YouTube comments. Our text data is the 21.8-million-word Sexism subcorpus of the NetLang corpus of YouTube comments. In order to identify instances where YouTube commenters explicitly suggest their personal involvement in carrying out physical violence, we analyse aggression verbs in constructions with the first-person singular subject pronoun (e.g., *I slap, I kill*). Through manual classification of the instances based on their target and function, we lay out a taxonomy of how aggression verbs are used in online comments and compare the various functions associated with the studied aggression verbs. We find that the pattern with the first-person singular subject and aggression verb frequently indicates the commenter's subjective aggressive perspective towards the target. The aggression verb constructions are used in threats and threatening wishes targeting specific individuals rather than vulnerable groups, and threatening language contributes to the online norm of aggressive language.

Keywords. Verbal aggression, hate speech, online hate, first person, YouTube, sexism, threats, targets of verbal aggression, functions of verbal aggression, aggression verbs

List of concepts for index. Verbal aggression, hate speech, online hate, first person, YouTube, sexism, threats, targets of verbal aggression, functions of verbal aggression, aggression verbs, coercion, automatic detection of hate speech, impoliteness, YouTube's hate speech policy, protected groups, incitement, keyword analysis, GloWbE Corpus, inductive approach, clustering

1 Introduction

Verbal aggression is regarded as one of the downsides of the participatory culture of online social media platforms as it has both individual and societal repercussions and may contribute to harmful and criminal phenomena such as cyberbullying, hate speech, cyberterrorism and even war (e.g., Bilewicz and Soral 2020, Chetty and Alathur 2018, Hamilton 2012). Especially sites allowing users to comment on posts by other users anonymously, such as YouTube, are often described as environments prone to aggressive and derogatory communication (e.g., Brown 2018). Earlier research on hateful language on these sites has focused, for instance, on individual words and lexical co-occurrence patterns that could be used to detect hate speech automatically (see, e.g., Tontodimamma et al. 2021), but our interest in online verbal aggression stems from a somewhat different research perspective.

We are interested in how the commenters position themselves in this type of online interaction and ask whether and how they attribute aggression to themselves in the first person. The first person in general is an important locus of interpersonal and indexical work and it often establishes the starting point of the interaction (e.g., Agha 2007: 280, Mühlhäusler and Harré 1990). The first-person singular attributes

maximum overt responsibility for what is being said to the self, and how we thus position ourselves in interaction may vary according to many situational factors and the types of roles we may assume in different social contexts. Verbal or physical aggression is not socially acceptable on most occasions, but forms of verbal aggression seem to occur regularly online (Hamilton 2012).

Our approach combines exploratory corpus-based investigation of patterns of 'I + aggression verb' with inductive qualitative close-reading of aggressive comments in context. Rather than explore turn-taking patterns in specific comment chains and the commenters' local positions in such discussions, we make use of the big data nature of the corpus to identify a more general taxonomy of first-person singular aggression online. To achieve this, we investigate the textual functions of aggressive comments (cf. Mahlberg 2007).

The data include YouTube comment sections relating to sexism collected in the NetLang corpus. YouTube is the biggest global online video platform where videos are uploaded both by organizations and private individuals. The data source of this study is the YouTube comment section: any logged-in user can leave a comment on a video, and the comments will be visible to all other users. (For an overview of YouTube, see, e.g., Androutsopoulos and Tereick 2015). In this paper, the extracts from the data are quoted verbatim; while we have avoided including instances of highly graphic descriptions or threats, the analysis will by necessity include quotes containing aggressive, sexist and misogynistic language, including references to sexual assault.

2 Background

The volume of research on hate speech has increased rapidly during the past fifteen years in tandem with the development of online participatory media, especially in the fields of automatic hate speech detection, legal scholarship and social sciences (Tontodimamma et al. 2021). Language scholars have been slower to adopt the hate speech framework, possibly due to its restrictive definition. In many studies exploring some form of derogatory or offensive language use, scholars are not thus restricted but deal, for instance, with online bullying (McCambridge 2022), or they approach hate speech with a pragma-linguistic framework such as impoliteness (Culpeper 2021, Culpeper et al. 2017). In our analysis, we adopt a broader, less limiting term *verbal aggression* to refer to "the act of using aggressive language on a target" (Hamilton 2012: 6). In this way, we cover a broader territory of first-person aggressive verbal behaviour than concentrating strictly on hate speech would allow us to do. As far as we know, there are no previous studies that would specifically target verbal aggression in first-person singular verbal constructions.

While there is no universal definition of hate speech, it is generally defined as "bias-motivated, hostile and malicious language targeted at a person or group based on their actual or perceived characteristics", such as ethnicity, political orientation, or gender (Siegel 2020: 57). That is, hate speech is directed at an individual due to their (assumed) membership in a group (Sellars 2016: 25). In legal definitions "incitement to discriminatory hatred" is another key factor in addition to the vulnerable characteristics (Culpeper 2021: 5, see also Baider 2022, for a discussion of definitions). For example, YouTube's hate speech policy describes hate speech as promoting violence or hatred against individuals or groups, based on their age; caste; disability; nationality; race or ethnicity; sex, gender identity, or gender expression; or sexual orientation, among other features (YouTube Help n.d.: a). Some of the examples in the dataset analysed here match this strict definition of hate speech. Yet, verbal aggression, threats and incitements of violence can also be attributed to the victim's personal or situational factors, such as their appearance, perceived authenticity or likeability (McCambridge 2022), or behaviour, such as Verbal Trigger Events (see Wigley 2010). In addition, the danger posed by antagonistic language ranges from offense and harassment to calls of discrimination and even to coordinated extremism and attacks (Gagliardone et al. 2016: 18–19).

The linguistic forms of verbal aggression cover many types of hostility, rudeness and incivility (cf. types of conventionalized impoliteness, Culpeper 2011: 135–136). Based on research on discrimination and racism, we know that such discourse may employ an array of discursive strategies including metaphors and

derogatory naming practices to construct the self/in-group as good and the other/out-group as bad (Reisigl and Wodak 2001). For example, recent studies using machine learning methods have identified words such as *fuck*, *ass*, *shit*, *faggot* and *little* among the top words in both offensive and hateful tweets, and additionally words like *hate* and *kill* in hateful tweets (Watanabe et al. 2018). However, since our analysis focuses on the first-person *I*, further tied to the presence of an aggression verb, many common types of verbal aggression are excluded from consideration, including various types of insults, such as personalized negative vocatives, assertions and references expressed in the second person (e.g., *you idiot*, *you are X*), as well as silencers (e.g., *shut the fuck up*) and negative expressives (e.g., *go to hell*, *fuck you*) (Culpeper 2011: 135–136).

Online spaces may be particularly prone to induce hateful content. The anonymity or pseudonymity of many social media discussions, including YouTube comments, has been associated with hostile behaviour for self and social identities. Users' sense of anonymity in online discussions may disinhibit individuals to engage in behaviour they would avoid in offline interaction (Suler 2004). Despite YouTube's policy prohibiting threats and "implied calls for violence" on the platform (YouTube Help n.d.: a, c), we found plenty of aggressive comments in the data, especially different types of *threats*, which we view as a particular type of verbal aggression. Threats can be formally defined as illocutionary speech acts that express the speaker's intention to commit an act that will affect the addressee negatively (Fraser 1998). However, the speech act does not entail an actual commitment to the act, only the intention is expressed (Fraser 1998, see also discussion of intentionality by Culpeper 2011: 48–52). Since we found many other types of aggressive comments, which do not seem to serve specific speech acts, we focus our analysis broadly on textual functions of verbal aggression instead (cf. Mahlberg 2007).

Hate speech and verbal aggression may serve many general functions. For example, applying the analytic framework of conventionalized impoliteness formulae, Culpeper et al. (2017: 14) found out that the language of religiously aggravated hate crimes is often coercive, seeking to force a realignment of values between the producer and the target by means of insulting and threatening the target. Online hate may also function to discourage targeted groups from participation (e.g., Nadim and Fladmoe 2021, Richardson-Self 2021) by inciting negative emotional reactions among the targets (e.g., Staude-Müller et al. 2012) and thus creating an unsafe atmosphere. Online hate is even known to incite extremism and real-life violence (e.g., Siegel 2020, Richardson-Self 2021). For example, members of the online 'incel' community, known for their hateful orientation towards women, have committed several violent acts in recent years (Pelzer et al. 2021).

From the perspective of the instigator, there may be a number of different motivations to engage in online hate. For example, some individuals engage in hate speech for a political cause or as a part of a cultural struggle, others use it as a way to draw attention to social problems (Erjavec and Kovačič 2012). While online hate can be purposeful and ideological, it can also be affectively motivated, spontaneous and related to the individual's subjective emotions (Saresma et al. 2022: 89). For example, online aggression may be triggered by the speaker's sense of being threatened or being put in a weaker position by an out-group (Saresma et al. 2022: 93–94). A sense of anonymity in a crowd of other users can further encourage users to see others as part of stereotyped out-groups (Postmes et al. 1998, Spears and Postmes 2015). Impolite or hostile comments may boost in-group identity by creating affiliation among commenters. For example, criticising an out-group or a member thereof may strengthen a sense of in-group homophily among YouTube commenters (Andersson 2021), with the in-group judgement co-constructed as a mass social judgement attributed to "everyone" (e.g., McCambridge 2022).

Similarly, although hate speech and online hate are often linked with societal issues and political ideology, for many, producing hateful content is simply "just some game in the online community" (Erjavec and Kovačič 2012: 912). So-called performatively motivated hate speakers may share opinions or ideas that they do not truly support in order to provoke or escalate an argument and get a reaction (Saresma et al. 2022: 90). This type of action can be described as trolling, i.e., "deliberate, deceptive and mischievous attempts that are

engineered to elicit a reaction from the target(s) [and] are performed for the benefit of the troll(s) and their followers” (Golf-Papez and Veer 2017: 1339). For example, some Twitter posts containing sexual aggression may contain threats and harassment to cause fear in the recipient, whereas others contain no linguistically evident intent to harm and are instead posted as part of jocular discourse of spam or ridicule (Hardaker and McGlashan 2016). Nevertheless, since the motivation is not known to the target, even online hate with performative or jocular function can cause emotional distress for the affected individuals and groups. Indeed, while some definitions of hate speech consider whether hate speech causes negative effects beyond the speech itself, these are problematic precisely because of the challenges involved in measuring the harm to the victim(s) or the spread of hatred among readers (Sellars 2016: 27, Hietanen and Eddebo 2022).

Considering that online hate may often be motivated by ideological forces, it is not surprising that YouTube videos on controversial topics tend to attract more spam-like or offensive comments, emotional shout-outs and irrelevant content than do videos on more neutral topics such as animals (Schultes et al. 2013: 666). Because of the above reasons and because the original corpus compilation was based on the presence of words of verbal aggression, we assume that the data on a debatable topic such as gender contain offensive comments relating to (but not necessarily limited to) sexism. Indeed, our analysis demonstrates that YouTube tends to contain a noticeable amount of insulting, threatening and aggressive comments, despite YouTube’s hate speech policy, user’s ability to report hate speech and comment sections, and YouTube’s comment moderation targeting hate speech and cyberbullying (YouTube Help n.d.: a–c).

3 Methods

We looked at the most common aggression verbs occurring in the NetLang Sexism subcorpus, hypothesizing that some verbs would be used to express hate in constructions with the first-person singular. The corpus contains 21.8 million words representing over 762,000 comments posted in response to 172 different YouTube videos. The data includes comment sections of gender-related YouTube videos. The data was scraped automatically by the NetLang team using specific key words selected by the team to identify verbal aggression. The key words used to identify comment sections containing sexism included the following: 'Male chauvinism', 'Chauvinist', 'Gender', 'Sex', 'Sexual', 'Sexism', 'Misogyn', 'Misogyny', 'Misogynous', 'Misogynist', 'Misogyne', 'Patriarchy', 'Pussy pass', 'Misandry', 'Misandris', 'Woman', 'Chick', 'Dame', 'Old hag', 'Hag', 'Crone', 'Witch', 'Minger', 'B*tch', 'Froggie', 'Harlot', 'Ho', 'Hooker', 'Promiscuous', 'Sharmoota', 'Slag', 'Slapper', 'Slattern', 'Sl*t', 'Sloot', 'Tart', 'THOT', 'Trollop', 'Tramp', 'Wh*re', 'Dumb blonde', 'Becky', 'Make me a sandwich', 'Bimbo', 'Feminazi' (see Henriques et al. 2019). As the corpus comprises the identified comment sections in their entirety, it includes also comments without any verbal aggression or with verbal aggression relating to issues other than sexism.

First, the 500 most frequent verbs in the corpus were manually scanned to identify potentially ‘hateful’ verbs: the list was reviewed several times after which concordances of each verb were carefully checked to ensure they were used for verbal aggression at least some of the time. The following verbs were chosen for further qualitative inspection: *die, kill, fuck, shut, destroy, beat, break, throw, rape, slap, kick, punch, shoot, burn, smack, and rip*.¹ In an exploratory pre-analysis, we used keyword analysis (e.g., Scott 2010) to compare the frequencies of words in the study corpus and in the US-subcorpus of the GloWbE corpus (Davies and Fuchs 2015); this confirmed that most of the chosen verbs are particularly frequent in the study corpus and might thus reflect language characteristic to (sexism-related) verbal aggression in particular.

¹ Verbs that were initially inspected but were left out of the analysis are: *fight, blame, hit, hurt, fall, force, abuse, cut, attack, harm, punish, screw, suck, murder and bang*. With some of these there were isolated instances of hateful use, excluded from the analysis due to insufficient frequency.

For each verb selected for further analysis, we extracted concordances where the pronoun *I* appeared within 5L of the present tense form of the verb. The 5L scope allowed for modal verbs as well as intensifiers or other adverbials to appear between the verb and the pronoun, which was commonly the case. In the concordance searches, we included the base form of the verb, the third-person present tense -s form and present progressive -ing form but excluded past tense (see below). This initial selection resulted in 6,363 concordances (see Table 1). Next, we proceeded with a manual inspection of the concordances to ensure that they fit our final selection criteria.

The first criterion was that the concordance needed to include clear verbal aggression, for example, expressing violent actions towards the target as a function of the verb. Certainly, this part of the process hinged on the researchers' interpretation, as understandings of what is considered aggressive or threatening may differ between cultures and individuals (cf. Culpeper 2011: 14–15). Second, we confirmed manually that the aggression verb was directly attributed to the imagined actions or thoughts of the agent behind *I*. Third, we only included instances where the aggression targeted people, people-related abstract entities (e.g., humanity, feminism), and in some rare cases inanimate objects (such as computer screens). Following these criteria, we excluded concordances with non-aggressive meanings of the verbs as well as concordances with unfitting targets (e.g., *dying of laughter*, *beating someone in a game*, *destroying someone's arguments*).² While we made a note of the type of target of the verbal aggression, we did not specify the gender of the target or whether the aggression was triggered by sexism. Since we were interested in verbal aggression that occurred in the communicative event itself, we also excluded instances where the writer was describing a past event and/or someone else's actions (e.g., *I've seen a woman punch a complete stranger*), cases of reported speech (e.g., *I'd be like "ok I'd punch you"*),³ and meta-level discussion of violence (e.g., *I deserve to punch women as hard as I punch men*); Similarly, we excluded humorously-intended or sarcastic comments (e.g., *I use to beat my four wives just for fun everyday*), since such comments do not represent the writer's aggressive position in the same way as our other examples.

This process resulted in a final selection of 1,591 concordances (25% of the initial selection, see Table 1), with great variation between the different verbs. Whereas 70% or more of the instances with *punch*, *slap*, and *smack* were identified as verbal aggression, included in the analysis, only a small fraction of concordances with *rape*, *break*, and *fuck* were included. For example, many instances of *rape* turned out to be cases of reported speech, whereas *break* and *fuck* were utilized in many non-aggressive expressions; *fuck* mostly appeared in other roles than a verb.

² Similarly, with one relevant exception (*throw up*, see section 4), we excluded phrasal verbs such as *fuck up* and *kick out*.

³ Instances of past tense are thus mostly excluded. A rare exception are cases where the writer clearly talks about their reaction to the video in past tense; these are equivalent to other writers expressing their reactions in present tense, e.g. *I wanted to punch my screen*.

Table 1. Distribution of initial items and verbs included in the analysis

	item total	verb included	verb included %
beat	646	323	50 %
break	263	22	8 %
burn	92	23	25 %
destroy	77	18	23 %
die	801	126	16 %
fuck	2380	50	2 %
kick	163	73	45 %
kill	476	131	28 %
punch	337	236	70 %
rape	212	13	6 %
rip	80	13	16 %
shoot	96	32	33 %
slap	409	343	84 %
smack	158	122	77 %
throw	273	66	24 %
total	6463	1591	25 %

With an inductive approach (see, e.g., Thomas 2006), the analysis aimed at identifying the targets and functions of verbal aggression. With functions, we refer broadly to often-localized textual functions (cf. Mahlberg 2007). During the first rounds of analysis, the concordances of each verb were analysed separately. This allowed us to pay particular attention to verb-specific uses and functions. Approaching the data with a close reading, we coded the type of target and function of the verbal aggression for each concordance. The task was carried out by the first and second authors, who each annotated the concordances of about half of the selected verbs, discussing coding procedures and initial codes regularly. After the first round of coding, the authors decided on a final coding scheme. Each author then revised the coding of their share of the concordances. To further improve inter-annotator agreement, as the next step, each author inspected the concordances initially analysed by the other author, making note of any disagreement in the coding procedure. Deviant cases were discussed together, reaching a mutual decision in all cases. Last, the concordances of different verbs were organized in one file, allowing us to carry out further analyses and easily compare different verbs, targets and functions.

The targets of verbal aggression were categorized based on the specificity of reference. Specific targets include individual persons, groups and the self, whereas unspecific targets include generic references (e.g., *anyone, a woman*) and abstract references (e.g., *sexism, Islam, humanity*). While the analysis of functions was largely inductive and we approached the data with no particular framework in mind, once we had finalized the coding scheme, it was evident that *threats* were a common nominator for many of the identified functions (cf. Fraser 1998). As such, at the first level of our taxonomy, we categorize functions as ***threats of physical aggression*** and as ***expressions of mental aggression***. Threats were further classified as either *simple* or *conditional* threats, whereas we distinguished between *boulomaic* and *emotive* expressions of mental aggression. Third-level categories reflect our coding scheme and are further presented in section 4.1.

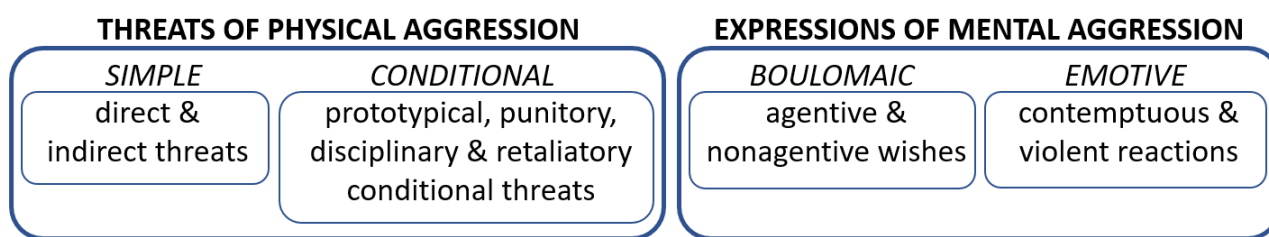


Figure 1 Taxonomy of first-person verbal aggression

Importantly, in our analysis of functions, we only consider the immediate textual context of the verbal aggression (L10–R09). At times, a broader context including the YouTube video might, for example, reveal relevant information about the nature of the comment, but multimodal analysis of this type falls beyond the scope of this chapter. In addition, while some overlap occurs between the function categories, we have opted to categorize the threats only based on their most salient function.

In section 4, we present verbatim examples from the data. The YouTube comment section is available to anyone without registration, and thus, there is no reasonable expectation of privacy for the commentators (European Commission 2021: 13–14). Nevertheless, following Scott (2022: 157), we considered whether the data excerpts could lead to the identification of the commentators. However, the examples have low searchability and thus prevent identification of the original contributions.

4 Analysis

We begin this section by introducing the functions represented by our taxonomy in 4.1., followed by a quantitative consideration of the association between verbs and functions in section 4.2. In section 4.3, we then examine the targets of verbal aggression, before discussing the results in more depth in section 5.

4.1 Functions of first-person verbal aggression

4.1.1 Threats of physical aggression

This main category includes both simple threats (*direct* and *indirect threats*) and conditional threats (*prototypical*, *punitive*, *disciplinary*, and *retaliatory*), the former lacking the specificity and detail as to why or under which condition the threat ought to be carried out, present in the latter.

4.1.1.1 Simple threats

4.1.1.1.1 Direct threats — *I'd kill that shrimp*

As the most common type of threats, with 355 instances, direct threats express the writer's (alleged) intention to harm the target of the threat, typically in a straight-forward fashion. Yet, while many direct threats are succinct (1–2), some include additional details concerning the way the threat might be delivered (3); detailed accounts were particularly frequent with *kill* (4).

(1) *I hate you I will kill you*

(2) *I'll fucking destroy you*

(3) *I will bitch slap you with a hammer until you WAKE UP*

(4) *I would fucking kill her I would shoot her in the legs beat the shit out of her with a crow bar*

Direct threats were often expressed with future orientation, for example with the auxiliary verb *will* (1–3). However, over half of direct threats (196 out of 355 instances, or 55%) appeared in *would* constructions (4–6). While the modal verb *would* suggests conditional mood, we categorized these constructions as simple

(and direct) threats when *would* appeared alone and the condition under which the threat would be acted upon was not specified, in contrast to *conditional threats* (see below).

(5) *I would kill that racist bitch*

(6) *I'd beat tf out that bitch like dead ass*

4.1.1.1.2 *Indirect threats — Good thing I know how to break a wrist with my bare hands*

Indirect threats are a function category distinct from other types of threats. The category was not frequent in the data (with only 22 instances) and thus no salient constructions were identified with indirect threats. Instead, indirectness emerges at the level of meaning. Indirect threats do not declare the writer's intention to act against the target, but they suggest the possibility of violence, for example by stating that the writer has the ability to carry out an act of aggression (7) (see Yamanaka 1995: 52). Other times, indirectness rose from the vague level of intent (8) or in the form of a question, for example (9).

(7) *oi mate bet I could fuckin beat your ass*

(8) *I had the thought of kicking those 3 girls*

(9) *do i punch you or punch you??*

4.1.1.2 *Conditional threats*

Conditional threats contain the writer's threat of committing an act as well as a specification of a condition: a reason for the act to occur, a prerequisite or hypothetical situation in which the act would be carried out. Based on the type of conditional, we categorize threats as prototypical, punitive, retaliatory, and disciplinary. As indicated by the category label, we classify typical cases as prototypical conditional threats, which lack further qualities found in punitive, retaliatory and disciplinary conditional threats.

4.1.1.2.1 *Prototypical — If I was there, I would slap her*

Prototypical conditional threats were nearly always expressed with the construction *if X, then I would Y*. The conditional *if*-clause suggests that the threat would be carried out under some hypothetical circumstances. However, prototypical conditional threats resemble the simple threat function category in that the writer does not specify an obvious reason as to why the act ought to be carried out, as instead the condition is given along the lines of "if I was given the chance" or "if it were me" (10–12).

(10) *If I see this girl out in public I will slap the living daylight out of her*

(11) *If I was god I would first destroy the feminists, wouldn't think a second*

(12) *If I were you I would throw her across street and leave her there*

Details on the reason or condition may be evident to readers if the actions of the addressee are given in the YouTube video, as is the case in example (10), where the target, "this girl", is a person featured on the video.

4.1.1.2.2 *Retaliatory — If someone grabs me, I'll kick their ass*

A special type of a conditional situation occurring in the data was that of retaliatory conditional threats. Similar to prototypical conditional threats, retaliatory threats were often expressed with the construction *if A does X to me, then I would Y*, hence specifying the action that would lead to the retaliatory threat being carried out. For example, retaliation was imagined to happen if someone attacked the writer first (13), if someone misbehaved towards the writer (14) or made racist remarks, for example (15).

(13) *if they punch i [sic] punch them back*

(14) *If she ever talked to me like that I would beat her ass*

(15) *If somebody white call me a [n-word] I would kill then [sic]*

Retaliatory threats seem to be utilized to present the aggression as “reasonable” or “justified”. This way, the writer takes a stance towards actions they see as unacceptable while perhaps also giving a warning to any addressee that might initiate conflict.

4.1.1.2.3 *Punitory — I still punch a toxic woman in their face*

Punitory conditional threats were usually also expressed in similar ways to direct threats, with the distinction that with punishments, the writer clearly indicated a reason for the violent act to occur. While retaliatory threats indicate that aggression would occur in response to a potential action of an oftentimes generic target, punitory threats generally occur in reaction to the target’s alleged behaviour (16–17), prejudiced attitude (18–19), perceived stupidity or intellectual capacity (20), or ideological positioning (21). At times punishments were explicitly framed as conditional, with the modal verb *would* (16–17, 20).

(16) *I would slap if she gave me that fuckin attitude*

(17) *This girl is so cringy I would beat her into a 3yr coma*

(18) *I’ll punch her so badly stupid racist ugly bitch*

(19) *I won’t hesitate to punch that sexist son of a bitch*

(20) *I’d end up slapping some idiots with the good lords bible*

(21) *I’ll freaking kill u because ur a feminist u offend me*

Commonly, the punishable behaviour or characteristic of the target on the video was referred to implicitly. By threatening punishment, the writer judges the target’s behaviour or attitude as unfavourable or unacceptable some way or another.

4.1.1.2.4 *Disciplinary— I would have to beat her ass if I were her parents*

Sharing characteristics with both prototypical conditional threats and punitory threats, disciplinary conditional threats include specific threats that suggest corporal punishment in order to teach the target some kind of a lesson. In these cases, the hypothetical condition is used by the writer to situate themselves in an authoritative role, typically as the target’s parent (22–23) but sometimes also as a sibling (24).

(22) *If she was my daughter I will smack the shit out of her*

(23) *If I was her mama ohh I swear I’d slap and spank her everyday until she learns her lesson*

(24) *If I was her sister I would rip her head off*

Because of the prevalence of parental punishment, we included in the disciplinary category instances where an image of a parent punishing a child is otherwise evoked, with mentions of lashing someone with a slipper or a belt (25), for example.

(25) *Her bitchy ass needs the belt and a good slap of reality.*

On the other hand, implications of punishing someone physically in order to teach them a lesson were also attested in uses of the verb *fuck* (26), where the writer situates themselves in a position of power and control over the target but not as an imagined family member.

(26) *I wanna fuck the racism out of her*

4.1.2 Expressions of mental aggression

The data also includes types of verbal aggression that lack the indication of the writer's intention to act, which is associated with threats (see Fraser 1995). Instead, they convey internal mental states that the first-person writer ascribes to themselves. This main category includes boulomaic expressions conveying a wish and emotive expressions evaluating the target.

4.1.2.1 Boulomaic expressions

Boulomaic expressions were used to express what the writer hopes will happen to the target of verbal aggression; in this sense, these expressions are characterized by a threatful quality. Boulomaic expressions were further distinguished based on whether the writer wishes the action were to be carried out by themselves (*agentive wish*) or by someone else (*nonagentive wish*). Particularly agentive wishes resemble *threats* in that an aggression verb indicates violence towards a target, and the *I* functions as an imagined agent. However, framing the act of aggression as something the writer allegedly merely desires to do or hopes to happen mitigates the writer's intention to carry out the act of aggression.

4.1.2.1.1 Agentive wishes — *I wanna slap her so hard*

Agentive wishes were most typically expressed with constructions using boulomaic verbs, for example *I hope*, *I want to* (or *I wanna*) and *I wish* (27–29). Notably, the construction *I want to [beat/kick/slap/smack] the [shit/crap/fuck] out of [target]* was utilized fairly frequently (examples 29–30). A less frequent but salient construction was asking for permission to do something, sometimes even followed with a polite *please* (30–31).⁴

(27) *I wanna punch this bitch in the throat so badly*

(28) *My head hurts I just wanna kill feminists that are like this*

(29) *I wanna slap the shit out of her*

(30) *Can I kick the shit out of her soul?*

(31) *Can I just slap her, please*

Importantly, agentive boulomaic expressions imply that the writer does not intend to carry out an act of violence. Although this may, of course, be because the writer is unable to do so, we see this category as mitigating the intention of the writer compared to the intention entailed by a direct threat.

4.1.2.1.2 Nonagentive wishes — *I hope she dies*

Nonagentive wishes represent the writer's wishes, but the writer is not positioned as the agent of the desired action.⁵ Most commonly, there was no named agent for the action (32–34), but sometimes there was an unspecific agent, e.g. *someone* (35). Only rarely were agents of these wishes specific people (36). Apart from missing an agent, nonagentive wishes were formulated in similar ways as agentive wishes (e.g., *I hope*, *I wish*).

(32) *i hope she kills herself and saves the world a lot of trouble*

(33) *I think he should burn in hell*

(34) *I wish all racists would die not just white*

⁴ While these constructions were common for wishes, they were not exclusively used in boulomaic expressions, for example, *I want to throw up* was classified as a contemptuous reaction (see below).

⁵ Since our selection criteria excluded past tenses of the selected verbs, many common types of nonagentive wishes expressed in passive constructions are thus excluded, e.g., *I hope she gets punched*. Hence, this category does not represent all nonagentive wishes.

(35) [...] *and i hope someone kills your husband*

(36) *I hope he fucks you till your [sic] dead*

The verb *die* was frequently used in this category, given that this intransitive verb cannot have an active agent who would carry out an act towards a target. On the other hand, nonagentive wishes of the target's death were attested with the verb *kill* (32, 35) and other lexicogrammar detailing circumstances, such as *until you're dead* (36).

4.1.2.2 *Emotive expressions*

The last category of emotive expressions differs from all other categories in that the (imagined) physical violence or act of aggression does not directly target the cause of the aggression. Instead, the physical aggression verb is directed towards the writer's self or towards inanimate objects (*violent reactions*), or without directing the act at anyone or anything (*contemptuous reactions*). However, we chose to include emotive expressions in our analysis, since they express hostility and intense dislike towards a target, which qualifies as verbal aggression. That is, the verbal aggression is typically triggered by someone or something in the video. This trigger of aggression is the ultimate target of the negative evaluation by the writer, thus tying the emotive expressions indirectly to similar targets as we have included above.

4.1.2.2.1 *Violent reactions — I wanna kill myself after this video*

Violent reactions were typically explicit reactions to the video: the video, someone or something in the video, makes the person want to act aggressively (37–40). The verbs *die*, *kill* and *punch* were most frequently employed in this function. *Die* and *kill*, used in a hyperbolic way, typically directed the action towards oneself (37–38), while inanimate objects such as the wall or a computer screen were found with *punch* (40).

(37) *I just want to die because I watched this video*

(38) *I will kill myself before I go into a nursing home*

(39) *I'd probably rather shoot myself than spend 5 second with them.*

(40) *I really wanted to punch my monitor as soon as this discussing woman opened her mouth*

4.1.2.2.2 *Contemptuous reactions — I would not fuck her for gold bars*

We also identified specific uses with the verbs *fuck*, *throw* and *throw up* that functioned as manifestations of verbal aggression, yet without the threat or wish of violence present in the other functions. These contemptuous reactions were used to express extreme contempt towards the target.

Similar to violent reactions, *throw up* was employed for negative evaluation, to express that the video or something in the video was revolting enough to cause the reaction (41–42). Much less frequently, similar contempt was expressed by indicating one would *throw* something concretely, for example, in the trash (43).

(41) *I was about to throw up while watching this*

(42) *Ewww when they made out I throw up*

(43) *I will throw Quran in the recycle bin [...]*

Meanwhile, with *fuck*, contemptuous reactions were mainly expressed by stating that one would *not* fuck someone, often coupled with additional evaluative language (44) or circumstances (45). What makes this construction distinct from most others included in the analysis is that instead of threatening to do something, the commentators here are conveying the opposite, with the negated threat qualifying as verbal aggression given the contemptuous evaluation intended. As such, female worthiness is tied to male desire, and contempt is expressed by considering a female unworthy of male attention.

(44) *I wouldn't fuck one of these nasty slags either*

(45) *I wouldn't fuck her with a stolen dick*

4.2 *Functions associated with verbs*

To explore which verbs resemble each other in terms of their functions in the data, we clustered the verbs based on the percentual frequency of their functions as verbal aggression (dendrogram in Table 2).⁶ The frequencies of the functions are used to assess the typical functions of verbal aggression in our data and the dispersion of verbs across different functions. Five concordances from the initial selection (n=1,591) did not match our taxonomy and are here excluded from the analysis.

The most frequent functions are direct threats and agentive wishes, which suggests that most of the verbal aggression in the data does not specify the reason for why the act of physical violence should be carried out. On the other hand, conditional threats outweigh simple threats, and retaliatory conditional threats are the third most frequent function category. Together, retaliatory, punitive and disciplinary threats account for 485 instances out of the total of 1,586, meaning that 31% of the instances of verbal aggression are asserted or implied to be justified because of the unacceptable action, demeanour, or characteristic of the target.

⁶ The clustering was done using the 'stats' and 'hclust' packages in R. The distance between the verbs was measured as Euclidean distance and the distances were clustered using UPGMA clustering method.

Table 2. Clustering of verbs and proportional frequencies of functions

	THREATS OF PHYSICAL AGGRESSION						EXPRESSIONS OF MENTAL AGGRESSION				Total
	SIMPLE		CONDITIONAL THREATS				BOULOMAIC		EMOTIVE		
	Direct threat	Indirect threat	Prototypical	Punitory	Disciplinary	Retaliatory	Agentive wish	Nonagentive wish	Violent reaction	Contemptuous	
destroy	22%	6%	17%	17%	0%	22%	6%	0%	11%	0%	18
break	24%	10%	10%	10%	0%	43%	0%	0%	5%	0%	21
kick	33%	4%	11%	5%	7%	21%	16%	1%	1%	0%	73
beat	30%	2%	5%	5%	19%	26%	10%	3%	1%	0%	322
rip	31%	0%	23%	0%	15%	15%	15%	0%	0%	0%	13
throw	15%	0%	19%	0%	4%	19%	12%	15%	0%	15%	26
shoot	38%	0%	6%	0%	0%	6%	22%	3%	25%	0%	32
kill	28%	1%	13%	4%	0%	12%	18%	4%	21%	1%	130
smack	22%	1%	1%	14%	20%	15%	28%	0%	0%	0%	122
slap	24%	1%	8%	12%	14%	6%	34%	0%	1%	0%	343
punch	18%	2%	5%	7%	2%	20%	42%	0%	5%	0%	236
rape	38%	0%	0%	23%	0%	8%	23%	8%	0%	0%	13
die	0%	0%	0%	0%	0%	0%	0%	82%	18%	0%	126
burn	13%	0%	0%	0%	0%	9%	0%	65%	13%	0%	23
throw up	0%	0%	0%	0%	0%	0%	0%	0%	0%	100%	38
fuck	18%	0%	0%	4%	14%	0%	4%	4%	0%	56%	50
Total	355	22	95	108	152	225	335	142	81	71	1586

Table 2 shows that there is a lot of variation in the distribution of functions among the selected verbs. The most frequent verbs – *slap*, *beat*, *punch* – are dispersed across several functions. Despite semantic similarity of these verbs as expressing concrete violence inflicted by hands, *beat* is associated more closely with direct threats and retaliatory threats, whereas *punch* and *slap* as well as *smack* occur most commonly with the boulomaic agentive wish function, where threat towards the target is mitigated. *Slap* and *smack* are also used for disciplinary threats, further indicating that the physical threat of these two verbs is relatively mild compared to *punch*, for which only 2% of instances are disciplinary threats. Meanwhile, *kill* and *shoot* occur at highly different frequencies in the data but bear semantic resemblance and are associated with direct threats, agentive wishes, and violent reactions. In summary, although the most frequent verbs are distributed across almost all the functions, they are associated more strongly with certain functions of verbal aggression. This may in turn partly explain the frequencies of the functions: for example, the frequency of *punch* may explain the frequency of agentive wishes, as *punch* accounts for 98 out of the 335 instances (29%) of this function category.

On the other hand, some of the studied verbs are limited to specific functions. While *burn* and *die* occur at different frequencies, both are used primarily in the boulomaic nonagentive wish function. As noted above, the intransitivity of *die* explains why as many as 82% of the instances of this verb are categorized as non-agentive and why 103 out of 142 (73%) instances of nonagentive wishes are indeed wishing for the target’s death. *Fuck* and *throw up* are in a cluster distinct from the other verbs in that they are used for contemptuous

reactions; in fact, *throw up* is used exclusively to convey contempt. The remaining verbs are generally too infrequent or too widely dispersed in our data for any reliable generalization.

4.3 Targets of first-person verbal aggression

While individuals were rarely named in the data, verbal aggression was by far most frequently targeted at specific persons (Table 3). As illustrated by many of the examples in section 4.1, these targets were typically represented by a pronoun, e.g., *her*, *him*, or *you*, interpreted to be a person in the video. Similarly, groups represent people appearing in the video, referenced to with the pronouns *they* and *you*, and with various types of NPs, such as *these idiots*, *these motherfuckers*. Groups also often refer to categories such as *women* and *feminists* (example 28) or *racists* (34). While we did not quantify this aspect, the majority of the targets in the data seem to be women (and/or feminists); this is likely explained by the corpus compilation procedure.

Some verbs target the writer’s self (Table 4), typically expressed either with intransitive (*I + verb*) or with transitive constructions (*I + verb + myself*). However, as discussed above and shown in Table 4, this target category was restricted to the violent and contemptuous reactions, which suggest that the writer is prepared to harm themselves, yet entail negative emotive reactions triggered by some other individual or object. The “other” target category is also mostly covered by the violent reaction function, as “other” targets mainly include animals or inanimate objects in expressions of anger, such as *punch my monitor* (example 40).

Meanwhile, generic targets were often represented by NPs, such as *a girl*, *that bitch*, but also by personal pronouns (*you*, *they*) and indefinite pronouns, such as *anyone*, *someone*. In contrast, abstract targets such as feminism, sexism, racism, religion, or humanity at large were infrequent.

Table 3. Verbs associated with targets of first-person verbal aggression

	person	group	self	generic	abstract	other	total
destroy	11	4	0	1	2	0	18
break	10	1	0	9	0	1	21
kick	54	2	1	16	0	0	73
beat	238	13	1	69	0	1	322
rip	10	0	0	3	0	0	13
throw	19	1	0	4	1	1	26
shoot	17	6	8	0	1	0	32
kill	70	20	27	9	1	3	130
smack	88	6	0	28	0	0	122
slap	298	23	2	20	0	0	343
punch	150	17	1	59	0	9	236
rape	9	3	0	1	0	0	13
die	73	20	23	0	10	0	126
burn	12	5	2	1	3	0	23
throw up	0	0	38	0	0	0	38
fuck	40	3	0	3	1	3	50
total	1099	124	103	223	19	18	1586

Table 4. Functions associated with targets of first-person verbal aggression

	person	group	self	generic	abstract	other	total
SIMPLE THREATS							
direct threat	295	18	0	39	2	1	355
indirect threat	10	6	0	5	0	1	22
CONDITIONAL THREATS							
prototypical	76	6	0	13	0	0	95
punitive	75	11	0	22	0	0	108
disciplinary	146	5	0	1	0	0	152
retaliatory	83	9	0	130	1	2	225
BOULOMAIC EXPRESSIONS							
agentive wish	285	42	0	6	1	1	335
nonagentive wish	106	24	0	1	11	0	142
EMOTIVE EXPRESSIONS							
violent reaction	1	2	64	2	2	10	81
contemptuous r.	22	1	39	4	2	3	71
total	1099	124	103	223	19	18	1586

Table 2 shows that there are some tendencies for specific functions to be directed at certain targets. Overall, specific persons were the most common target of verbal aggression in the data, but they stand out as particularly common for various types of threats (*direct threats, prototypical and disciplinary conditional threats*) and for boulomaic expressions (both *agentive* and *nonagentive*). Other types of targets are less frequent for these functions, although generic targets did appear with direct threats, and abstract targets (e.g., religion, feminism) almost exclusively occurred with nonagentive wishes. This may reflect how it is not feasible for an individual to (physically) harm an abstract target, leading writers to wish for a more abstract nature of damage instead with *die* or *burn*, which were commonly associated with nonagentive wishes. What is more, while group threats do not seem to be associated particularly strongly with any functions, they do appear somewhat more frequently with *boulomaic expressions*. This may also reflect the more abstract nature of wishes; concrete acts of violence are (imagined to be) directed at specific targets, whereas an unspecified death is easier to wish upon a group.

What is interesting is that while many of the conditional threats (prototypical, retaliatory, punitive, disciplinary) could have imagined, generic targets, only retaliatory threats stand out in this respect. Indeed, retaliatory threats were often built by imagining the premises in which someone else’s behaviour would cause oneself to act violently (e.g., example 15). In contrast, disciplinary threats are almost exclusively directed at specific people, that is, used to criticize the actions of a specific individual on the video.

5 Discussion and conclusion

In sum, our analysis demonstrated that first-person verbal aggression takes many forms, and that despite the strict hate speech policy, YouTube comments include a plethora of “implied calls for violence” and threats of physical violence (YouTube Help n.d.: a). We specifically explored a selection of verbs that we identified as being utilized at least some of the time in verbal aggression: *die, kill, fuck, shut, destroy, beat, break, throw, rape, slap, kick, punch, shoot, burn, smack, and rip*. Our qualitative analysis revealed that these verbs are indeed employed in verbal aggression, but verbal aggression is not their only possible function. Such multifunctionality poses difficulties for automatic hate speech detection, but our findings and the taxonomy introduced may potentially be used to improve the accuracy of automatic recognition.

While previous research (e.g., Culpeper 2011: 135–136) has proposed various frameworks and taxonomies for what we call verbal aggression, our specific focus on the first-person and on aggression verbs necessitated an inductive approach to the analysis. We categorized instances of verbal aggression as different types of *threats of physical aggression* and *threatening expressions of mental aggression*, the latter category including some specific types of comments expressing *negative emotive evaluation*. Overall, *direct threats* (e.g., *I will kill you*) were the most common type of verbal aggression in the data, followed closely by threatening *boulomaic agentive wishes* (e.g., *I want to punch her*) and less frequent but distinct nonagentive wishes (e.g., *I hope she dies*); different types of *conditional threats* were also frequently employed, specifying the condition under which the threat would be carried out (e.g., *if you do X, I will punch you*). The analysis confirmed that aggression expressed from the point of view of the writer in the first person occurs regularly online and writers do not shy away from taking responsibility for the most explicit threats of violence.

The analysis also revealed that the targets of verbal aggression were most commonly specific people (in 69% of instances), typically either other commentators or people appearing in the video. While hate speech specifically targets people based on their salient group membership such as (assumed) gender, ethnic background or political orientation (e.g., Sellars 2016: 25, Siegel 2020: 57, see also Silva et al. 2016), we found that at least on the surface level, verbal aggression was often triggered by someone's behaviour. Nevertheless, we also found comments where the target's background was specified (e.g., *I'd slap that Asian bitch*). The data moreover included repeated references to women and feminists, but we must attribute this to the data collection method and thus cannot regard it as a finding.

Furthermore, while we found many instances in which verbal aggression at least seemingly occurred in reaction to behaviour (e.g., punitory threats targeting individuals who are perceived to have broken social norms), this is not to say that the individual's assumed membership in a group does not play a role in triggering verbal aggression. Indeed, a multimodal exploration including the videos might have revealed more information about the targets. A future study could consider how the person(s) and the content of the video predict the types of aggression in the comments, for example whether the threats or ideological views expressed in the video are reflected in the comment section. Certainly, the content of the video plays some part, as in the current study, we observed threats directed at women featured in the videos and several discussions on the topic of equality, where threats were used in meta-level discussions on whether it is justified (for a man) to punch a woman (e.g., *If someone slaps me, girl or boy, I'm going to punch them*).

What we did find surprising is that while hate speech in particular is directed at disadvantaged groups, verbal aggression did not target only disadvantaged groups but also the assumed aggressors, since targets were at times explicitly described as *sexists* or *racists* (e.g., *Shut the fuck up before i slap white racist ass*). Indeed, commentators regularly attempted to justify the aggression by various means — most notably occurring in threats that we labelled as *disciplinary threats*, in which the writer takes an authoritative, often parental role, to teach a lesson to the target for breaking a social norm (cf. Culpeper et al. 2017: 16–17, who found the correcting of past injustice and setting things right as a frequent perceived moral justification for harm).

Our focus on verbal aggression by the first person already frames the communicative acts as personal in a sense. However, we found variation in the severity of verbal aggression. Direct threats imply an intention to carry out physical violence, representing severe cases, while boulomaic expressions, indicating wishes, can be justified to simply express the writer's desires. Emotive expressions of aggression do not direct violent behaviour towards others, yet they carry out what seems to be one of the base functions of verbal aggression: an indication that someone or something is detestable to the extent that they “deserve” to be (verbally) abused.

Furthermore, while the focus of this chapter did not allow us to delve deeper into the matter, we found that the commentators were using various ways to either (somewhat) soften the threat of violence — or intensify

it. For example, the frequent use of *would* in threats might function to downtone the meaning (*I will kill her* vs. *I would kill her*), and sometimes verbal aggression was expressed as a question or even as a “polite” request (e.g., *Can I just slap her, please*). In contrast, meanings were regularly intensified with cursing and profanities (e.g., *I would fucking kill her, I hope this b*itch dies in the worst way possible*), as well as by adding gruesome details of the intended violence (e.g., *I hope you die slowly in a train wreck*). The frequency of instances of writers expressing threats or wishes of serious physical harm could suggest the normalization of aggression and to some extent the semantic bleaching of aggressive verbs such as *die*, *kill*, and *shoot*.

Last, we found threats and aggression directed at groups or at abstract societal concepts to be rare in the data. Instead, at least on a surface level, verbal aggression triggered by the target’s behaviour or personal characteristics often seemed to be spontaneous and motivated by the writer’s emotional reaction rather than by ideological goals (see Saresma et al. 2022). While the verbal aggression identified here does not necessarily match definitions of hate speech as incitement to violence or discrimination against a group, it may nevertheless contribute to or reflect broader ideological debates. Gendered, sexist and especially misogynistic stereotypes are a normalized part of harassment and trolling in internet culture (e.g., Condis 2018, Lumsden and Morgan 2018), which means that many instances of verbal aggression found in our data may to some extent draw on the broader discourse and contribute to normalizing the sexist discourse in online interactions.

References

- Agha, A. (2017). *Language and Social Relations*. Cambridge University Press.
- Andersson, M. (2021). The climate of climate change: Impoliteness as a hallmark of homophily in YouTube comment threads on Greta Thunberg’s environmental activism. *Journal of Pragmatics*, 178, 93–107. <https://doi.org/10.1016/j.pragma.2021.03.003>
- Androutsopoulos, J., & Tereick, J. (2015). YouTube: Language and discourse practices in participatory culture. In Georgakopoulou, A., & Spilioti, T. (eds) *The Routledge Handbook of Language and Digital Communication*. Accessed on 21 October 2022. <https://www.routledgehandbooks.com/doi/10.4324/9781315694344.ch22>
- Baider, F. (2022). Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. *International Journal of Semiot Law*. <https://doi.org/10.1007/s11196-022-09882-w>
- Bilewicz, M., & Soral, W. (2020). Hate speech epidemic: The dynamic effects of derogatory language on intergroup relations and political radicalization. *Advances in Political Psychology*, 41(1), 1–31. <https://doi.org/10.1111/pops.12670>
- Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, 18(3), 297–326. <https://doi.org/10.1177/1468796817709846>
- Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behaviour*, 40, 108–118.
- Condis, M. (2018). *Gaming Masculinity: Trolls, Fake Geeks, and the Gendered Battle for Online Culture*. University Of Iowa Press.
- Culpeper, J. (2021). Impoliteness and hate speech: Compare and contrast. *Journal of Pragmatics*, 179, 4–11. <https://doi.org/10.1016/j.pragma.2021.04.019>.
- Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence*. Cambridge University Press.
- Culpeper, J., Iganski, P., & Sweiry, A. (2017). Linguistic impoliteness and religiously aggravated hate crime in England and Wales. *Journal of Language Aggression and Conflict*, 5(1), 1–29. <https://doi.org/10.1075/jlac.5.1.01cul>
- Davies, M., & Fuchs, R. (2015). Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). *English World-Wide. A Journal of Varieties of English*, 36(1), 1–28. <https://doi.org/10.1075/eww.36.1.01dav>

- Erjavec, K. & Kovačič, M.P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6), 899–920, <https://doi.org/10.1080/15205436.2011.619679>
- European Commission (2021). Ethics and data protection. https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf (accessed 03 October 2022).
- Fraser, B. (1998). Threatening revisited. *Forensic Linguistics*, 5(2), 159–173.
- Gagliardone, I., Pohjonen, M., Zerai, A., Beyene, Z., Aynekulu, G., Bright., Bekalu, M. A., Seifu, M., Moges, M. A., Stremlau, N., Taflan, P., Gebrewolde, T. M., & Teferra, Z. M. (2016). *Mechacal: Online debates and elections in Ethiopia. From hate speech to engagement in social media*. Programme in Comparative Media Law and Policy. <https://eprints.soas.ac.uk/id/eprint/30572>
- Golf-Papez, M. & Veer, E. (2017). Don't feed the trolling: rethinking how online trolling is being defined and combated, *Journal of Marketing Management*, 33(15–16), 1336–1354
- Hamilton, M. A. (2012). Verbal aggression: Understanding the psychological antecedents and social consequences. *Journal of Language and Social Psychology* 31(1), 5–12. <https://doi.org/10.1177/0261927X11425032>
- 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>
- Henriques, P., Araújo, P., Ermida, I. & Dias, I. (2019). Scraping news sites and social networks for prejudice term analysis. In Weghorn, H. & Rodrigues, L. (eds.). *Proceedings of the 16th International Conference on APPLIED COMPUTING 2019*, pages 179–189, Cagliari, Italy, Nov 2019.
- Hietanen, M., & Eddebo, J. (2022). Towards a definition of hate speech—With a focus on online contexts. *Journal of Communication Inquiry*, Online preprint, <https://doi.org/10.1177/01968599221124309>
- Lumsden, K., & Morgan, H.M. (2018). Cyber-trolling as symbolic violence: Deconstructing gendered abuse online. In Lombard, N. (Ed.), *The Routledge Handbook of Gender and Violence*. Routledge
- Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social Science Computer Review*, 39(2), 245–258. <https://doi.org/10.1177/0894439319865518>
- McCambridge, L. (2022). Describing the voice of online bullying: An analysis of stance and voice type in YouTube comments. *Discourse, Context & Media*, 45, 100552. <https://doi.org/10.1016/j.dcm.2021.100552>
- Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora*, 2(1), 1–31. <https://doi.org/10.3366/cor.2007.2.1.1>
- Mühlhäusler, P. & Harré, R. (1990). *Pronouns and People: The Linguistic Construction of Social and Personal Identity*. Blackwell.
- Pelzer, B., Kaati, L., Cohen, K., & Fernquist, J. (2021). Toxic language in online incel communities. *SN Soc Sciences*, 1(231). <https://doi.org/10.1007/s43545-021-00220-8>
- Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries?: SIDE-effects of computer-mediated communication. *Communication Research*, 25(6), 689–715. <https://doi.org/10.1177/009365098025006006>
- Reisigl, M., and Wodak, R. (2001). *Discourse and Discrimination: Rhetorics of Racism and Anti-Semitism*. Routledge.
- Richardson-Self, L. (2021). *Hate Speech against Women Online: Concepts and Countermeasures*. Rowman & Littlefield.
- Saresma, T., Pöyhtäri, R., Knuutila, A., Kosonen, H., Juutinen, M., Haara, P., Tulonen, U., Nikunen, K. & Rauta, J. (2022). Verkkoviha: Vihapuheen tuottajien ja levittäjien verkostot, toimintamuodot ja motiivit [Online Hate: The networks, practices and motivations of the producers and distributors of hate speech]. Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja [Publications of the Government's analysis, assessment and research activities] 2022:48. <https://julkaisut.valtioneuvosto.fi/handle/10024/164244>
- Schultes, P., Dorner, V. & Lehner, F. (2013). Leave a comment! An in-depth analysis of user comments on YouTube. *Wirtschaftsinformatik Proceedings 2013*. 42. <http://aisel.aisnet.org/wi2013/42>.
- Scott, K. (2022). *Pragmatics Online*. Routledge.

- Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails.... In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 43–58). John Benjamins.
- Sellars, A. F. (2016). *Defining Hate Speech*. Berkman Klein Center Research Publication No. 2016–20. Boston Univ. School of Law. <https://doi.org/10.2139/ssrn.2882244>
- Siegel, A. (2020). Online hate speech. In Persily, N. & Tucker, J. (Eds.), *Social Media and Democracy: The State of the Field, Prospects and Reform* (pp. 56–88). Cambridge University Press.
- Silva, L., Mondal, M., Correa, D., Benevenuto, F. & Weber, I., (2016). Analyzing the targets of hate in online social media. *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, 687–690.
- Spears, R., & Postmes, T. (2015). Group identity, social influence, and collective action online: Extensions and applications of the SIDE model. In S. S. Sundar (Ed.), *The Handbook of the Psychology of Communication Technology* (pp. 23–46). John Wiley.
- Staupe-Müller F., Hansen, B. & Voss, M. (2012). How stressful is online victimization? Effects of victim's personality and properties of the incident, *European Journal of Developmental Psychology*, 9(2), 260–274. <https://doi.org/10.1080/17405629.2011.643170>
- Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, 7(3), 321–326. <https://doi.org/10.1089/1094931041291295>
- Thomas, D. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2), 237-246. <https://doi.org/10.1177/1098214005283748>
- Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: Topics of interest and their evolution. *Scientometrics* 126, 157-179. <https://doi.org/10.1007/s11192-020-03737-6>
- Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, 6, 13825-13835 <https://doi.org/10.1109/ACCESS.2018.2806394>.
- Wigley, C. (2010). Verbal Trigger Events — Other catalysts and precursors of aggression. In T. Avtgis & A. S. Rancer (eds.), *Arguments, Aggression and Conflict: New Direction in Theory and Research* (pp. 388–400). New York and London: Routledge.
- Yamanaka, N. (1995). On indirect threats. *International Journal for the Semiotics of Law*, 8, 37–52. <https://doi.org/10.1007/BF01677089>
- YouTube Help, n.d., a. Hate speech policy. <https://support.google.com/youtube/answer/2801939?hl=en>
- YouTube Help, n.d., b. Potentially inappropriate comments now automatically held for creators to review. <https://support.google.com/youtube/thread/8830320/potentially-inappropriate-comments-now-automatically-held-for-creators-to-review?hl=en>
- YouTube Help, n.d., c. Harassment & cyberbullying policies <https://support.google.com/youtube/answer/2802268?hl=en>