




Disruptive online communication: How asymmetric trolling-like response strategies steer conversation off the track

Henna Paakki^{1*} , Heidi Vepsäläinen²  & Antti Salovaara^{3,2} 

¹*Department of Computer Science, School of Science, Aalto University, P.O.Box 15400, FI-00076 AALTO, Finland (E-mail: henna.paakki@aalto.fi);* ²*Department of Computer Science, University of Helsinki, Helsinki, Finland;* ³*Department of Design, School of Arts, Design and Architecture, Aalto University, Espoo, Finland*

Accepted: 23 February 2021

Abstract. Internet trolling, a form of antisocial online behavior, is a serious problem plaguing social media. Skillful trolls can lure entire communities into degenerative and polarized discussions that continue endlessly. From analysis of data gathered in accordance with established classifications of trolling-like behavior, the paper presents a conversation analysis of trolling-like interaction strategies that disrupt online discussions. The authors argue that troll-like users exploit other users' desire for common grounding – i.e., joint maintenance of mutual understanding and seeking of conversational closure – by responding asymmetrically. Their responses to others deviate from expectations for typical paired actions in turn-taking. These asymmetries, described through examples of three such behaviors – ignoring, mismatching, and challenging – lead to dissatisfactory interactions, in that they subvert other users' desire for clarification and explanation of contra-normative social behavior. By avoiding clarifications, troll-like users easily capture unsuspecting users' attention and manage to prolong futile conversations interminably. Through the analysis, the paper connects trolling-like asymmetric response strategies with concrete data and addresses the implications of this nonconformist behavior for common grounding in social-media venues.

Keywords: Asymmetric responses, CMC, Common grounding, Conversation analysis, Disruptive communication, Trolling

1. Introduction

Trolling is one of the most problematic and most commonplace forms of antisocial online behavior. Because trolling is not always outright hostile, it is difficult for others to moderate or exclude from conversations. Trolls hide their true intentions (Donath 1999) by posting naïve, arrogant, dangerous, or incoherent comments (Hardaker 2013). If people engage by responding, the conversations may well turn

into proliferation of non-constructive comments – especially if the troll employs the right strategies to prolong the exchange.

Particularly alongside personal insults, such prolonged digression renders trolling detrimental to online conversation, not least because contra-normative behavior increases discontentment and disillusionment among contributors and other readers. While the literature has examined definitions of trolling (Coles and West, 2016; Hardaker 2010; Herring et al., 2002), the motivations or goals behind it (Fichman and Sanfilippo 2016, pp. 23–47; Sanfilippo et al., 2017b), the distinct styles and general trolling strategies manifested (Hardaker 2013; Herring et al., 2002; Lumsden and Morgan, 2017; on ideological trolling in particular, see Zelenkauskaite and Niezkoda, 2017), and opportunities for preventing such behavior (Golf-Papez and Veer, 2017), researchers have not studied the conversational or interaction strategies that prolong non-constructive conversations specifically in terms of conversation structure.

To address this gap, we studied trolling-like observable behaviors to discover precisely how troll-like users succeed in drawing others into taking part in futile discussions. We concentrated on carefully analyzing their conversation strategies at the level of turns in interaction. Our findings, from data that we gathered from several online forums, dealing with various discussion topics, revealed what appears to be a common denominator among successful disruption strategies: they make use of unconventional, asymmetric responses in conversational interaction. We present three prominent response strategies that troll-like users employ in overriding the social norms of conversation. Our findings have several important implications. The paper concludes by considering these, with special regard to whether anything can be done to prevent or detect possible trolling.

2. Theoretical background

The following subsections review prior research into the research phenomenon (trolling), the theoretical background we applied (common grounding, from psycholinguistics), and our methodological approach (conversation analysis).

2.1. Trolling

Identifiability levels in text-based online communication range from complete anonymity and pseudonymity to communication wherein one's true identity is known. Aspects of anonymity increase the potential for degenerative conversation involving harassment, hate speech, repetitive posts or spam, intensified polarization, and unresolvable disagreements (e.g., Coleman 2014). In addition, a lack of clues to authors' identity and, hence, their motivations and the intended tone of their posts complicates the interpretation of intent and meanings (Donath 1999; Sudweeks and Rafaeli, 1996). Trolls take advantage of these ambiguities to lure others into prolonged and fruitless discussion (Herring et al., 2002).

Belying its recent associations with politically motivated disinformation campaigns, trolling was originally mostly entertainment for many of those performing it (see Shachaf and Hara, 2010). In some overlap with this use, for such venues as some USENET groups, it served as a means of boundary regulation between in-group and out-group members, or group ‘veterans’ and novices (Bishop 2014; Graham 2019). When employed for other purposes, it still was aimed at causing harm to smaller groups of people rather than entire societies (Graham 2019; Phillips 2015; Tepper 1997). In contexts of more recent online discourse, in contrast, political influence seems to have emerged as a new key motivation for trolling (e.g., Badawy et al., 2018). This type of trolling is tuned to amplifying polarization, propagating political agitation, limiting freedom of speech, and spreading fear and disinformation (e.g., Akhtar and Morrison, 2019; Bennett and Livingston, 2018).

While trolling, especially in some of its older forms, might not always be harmful or malicious (Cruz et al., 2018; Kirman et al., 2012; Sanfilippo et al., 2017a), our interest stemmed from its detrimental effects. We found a good starting point in Hardaker’s research (Hardaker, 2010) into the effects of trolling on online communities, as well as common trolling styles (referred to as strategies by Hardaker 2013). Hardaker, who paved new ground for conversational research into trolling, defines trolls as discussion participants who feign sincerity while their real intention is to disrupt conversation or create conflict for the sake of amusement (2010, p. 237). She cites the following comment as an example: ‘Uh.....not to be nitpicky,,,,,but...the past tense of drag is dragged, not drug. Otherwise it was an interesting post’ (2013, p. 72). The troll manages to deceive others into believing that the post is sincerely written (on ‘pseudo-intentions,’ see Hardaker 2010, pp. 233–236). This post unnecessarily attacks another user’s ‘face’ in the discussion (see Brown and Levinson, 1987), hypocritically correcting the poster’s grammar although it contains errors of its own; thereby, it successfully provokes other users into retaliating and digressing from discussion of the original topic. While our attention is not confined to amusement-oriented trolling, Hardaker’s approach and classification of six perceived trolling styles form the foundations for our work.

She developed her classification by analyzing two large corpora of USENET discussion, from groups for horse-breeding- and football-related discussion. Filtering her search down to cases mentioning the word ‘troll’ or variants thereof, she derived the classification qualitatively. The trolling styles she identified range from overt, easily observable ones to covert (i.e., hidden or non-obvious) styles.

At the most overt end of her continuum is the ‘aggress’ category, which involves deliberately and openly provoking others into retaliating via aggression (e.g., name-calling or foul language), and shock trolling, which is characterized by posting about taboo or sensitive subjects such as religion, death, or touchy human-rights issues (cf. Phillips 2011; on ‘RIP trolling,’ see Fichman and Sanfilippo, 2016, p. 12). These trolling styles are very similar to ‘flaming’ (cf. ‘flame bait’ per Herring et al., 2002; Danet 2013; Herring 1994). Proceeding toward the more covert strategies, Hardaker defines ‘endanger’-class trolling as pseudo-naïve trolling (Donath 1999) wherein bad

advice is given under a cloak of innocence (cf. the ‘Outward Manifestations of Sincerity’ strategy described by Herring et al., 2002). More experienced users feel compelled to respond so as to protect the forum’s novices. In antipathy trolling, the troll creates a sensitive or antagonistic context to exploit by being purposefully provocative, covertly manipulating others in aims of eliciting emotional responses. A (hypo)critical troll usually criticizes others excessively for peripheral matters of form such as punctuation, quite possibly while producing spelling mistakes and the like in the critical posts. Finally, the ‘digress’ class of trolling is the most covert style on the continuum. It involves luring others into straying from the topic at hand by spamming, partaking in posting of (off-topic) material in the conversation thread, or introducing tangential topics (e.g., Herring et al., 2002; Hopkinson 2013; Turner et al., 2005). Although Hardaker gave names to all these trolling styles and presented dialogue excerpts to exemplify the various categories, her work focuses on general content-related (and some linguistic) features of trolling, not on closely analyzing trolling as an interactive process. We go further, with analysis that homes in on the details of how effective trolling unfolds at the level of turns in interaction.

2.2. Common grounding theory

While the studies cited above produced theory on trolls’ intentions and analyzed what they do, the researchers did not consider trolling in light of theories of ordinary social interaction. Our key premise is that ordinary well-intended communication manifests *common grounding*. This concept, developed in the psycholinguistics field, refers to a ‘collective process by which the participants try to reach a mutual belief that they have understood what each other meant’ (Clark and Brennan, 1991, p. 223). Accordingly, an unintelligible action, such as an unexpected turn in a conversation, calls for an explanation from its originator. This corresponds to the Gricean *cooperative principle*, under which all the parties conversing are expected to engage in rectifying breakdowns when suspecting misunderstanding or confusion (Grice 1975). The cooperative principle is expressed via four *maxims* that articulate preconditions for effective communication: the maxim of quality (what is said should be true), the maxim of quantity (what is said should be as informative as required), the maxim of relevance (what is said should be relevant), and the maxim of manner (what is said should be clear, unambiguous, brief, and orderly).

The central premise of common grounding theory is that everything that is said in a conversation needs to be grounded – i.e., turned into mutual knowledge. For reaching a shared belief that the content has been understood, the interlocutor must provide evidence, to maintain a perception of common ground in the interaction (Clark and Brennan, 1991; Clark and Schaefer, 1989; Krauss and Fussell, 1991). The effort expended on sustaining common grounding can be understood as a cost of communication. Building upon the Gricean cooperative principle, particularly its constituent maxims of quantity and manner, Clark and Wilkes-Gibbs (1986) have suggested that partners in communication seek to minimize the costs by applying a

principle of least collaborative effort: in conversation, the participants try to minimize their collaborative effort – the work that both do, from the initiation of each contribution to its mutual acceptance.

We argue that *trolling is in many ways antithetical to common grounding*. Communication with a troll-like counterpart incurs a higher cost than ordinary communication, because trolling-like behavior is designed to prolong regressive interaction and to cause communication breakdowns that others will attempt to mend. For example, troll-like users may express opinions without explaining them, or even without any apparent connection to the topic at hand. Efforts at trolling act contrary to the realization of common grounding by provoking others into co-creating patterns of regressive circular conversation that waste time and effort.

2.3. Conversation analysis

Many norms of conventional, non-technology-mediated conversation apply also to online conversations (e.g., Baron 2000; Herring 2010; Hård af Segerstad 2002). Hence, such phenomena as conversational repair, requests for explanation, and actions through which common grounding is achieved can likewise be found in online interpersonal interaction. So, naturally, trolls are happy to exploit them. For instance, arrogant responses such as unnecessarily addressing an earlier speaker as if talking to a child (Wagner et al., 2005) break the norms of common grounding.

Our study applied conversation analysis (CA) to examine interaction strategies that provoke others. As a research tradition that can inform careful analysis of interaction, CA builds on the analysis of social action in ethnomethodology (EM). Growing out of the EM tradition, which delves into micro-level social interactions, CA focuses particularly on analyses of turn-by-turn interactions between people. Both EM and CA study the methods by which humans coordinate their actions in social situations in a competent manner, how breakdowns in communication develop, and how they are resolved. Particularly relevant with regard to our interest in understanding trolling is the idea that social action depends on a moral obligation of behaving in an understandable manner. People who deviate from this norm are seen as breaking a fundamental rule. Typically, others challenge such deviators with demands that they should explain themselves (Heritage 1984).

CA therefore shares the view of psycholinguistic research on common grounding that discourse is jointly created in a conversation by its participants (Clark and Brennan, 1991; Schegloff et al., 1977). Both approaches apply micro-analytic attention to naturally occurring language in interaction. Our interest in CA is linked to its potential for analyzing specific actions that systematically break common grounding and prevent mutual understanding. The fact that we have not been able to identify prior studies of online trolling that have employed this method makes CA all the more interesting as a tool for analyzing online conversations.

Some characteristics specific to trolling must be taken into account in any CA-based analysis. It is an Internet-born phenomenon (cf. Hardaker 2017) that occurs

mainly in asynchronous online forums, in synchronous chat, and on other conversational social-media platforms. In online spaces, participants may often enter and exit a conversation without others' awareness. Also, text-based interaction is typically turn-based, with no temporally overlapping utterances. Although most conversation analysis studies do not deal with online interaction, some lines of analysis that focus on natively digital conversation are starting to emerge (Giles et al., 2015). These complement the extensive body of research into computer-mediated communication, or CMC (e.g., Dolata and Schwabe, 2019; Garcia and Jacobs, 1998; Herring 1994; Moore and Churchill, 2011; Moore et al., 2007; Ruhleder and Jordan, 2001), and EM-informed work in CMC (e.g., cf. Herring 2010; Rintel and Pittam, 1997; Woodruff and Aoki, 2004). CMC and EM-oriented studies have examined what makes online social interaction distinct – for instance, by pinpointing certain features that separate text-only online conversation from spoken interaction: turn-taking (e.g., Cech and Condon, 2004; Garcia and Jacobs, 1999), sequence organization, repair (Markman 2010), the norms related to responding (Skovholt and Svennevig, 2013), CMC's conversational maxims (Crystal 2001; Lindholm 2013), openings and lack of embodied conduct (e.g., Meredith 2019), etc. Scholars doing such work appreciate the fact that software-related factors may shape these and other aspects of the interaction, including the problems faced (e.g., misinterpretation of silences; see Garcia and Jacobs, 1999) – problems that may be exploited in trolling. Researchers have also pointed out similarities: e.g., participants in both spoken interaction and online interaction orient themselves to building sequentially organized courses of action and maintaining intersubjectivity (Meredith and Stokoe, 2014) – i.e., building and maintaining common ground.

3. Data and methods

Our approach to analyzing trolling-reminiscent conversation strategies was based on conversation analysis that focused on 'successful' trolling, cases in which a troll-like post gains several responses. Unsuccessful trolling-like behavior, such as toxic remarks that other parties ignore and do not react to, was beyond our study's scope; the latter type of behavior is not so detrimental to the discussion and to participants' attempts to maintain common grounding. In addition, only cases that feature responses are amenable to conversation analysis. That is, our focus on *successful* trolling-like behavior was in line with our focus on the impacts on a conversation and with methodological considerations. Thus, we were able to see how trolling-like behavior disrupted efforts to reach common grounding: while others in the conversation sought to maintain it, the trolling-like actions complicated this attempt.

The subsections that follow describe the creation of our corpus of trolling-like behavior and our methods for analyzing specific interactions between troll-like users and non-trolls. In Subsection 3.1, we present how we distinguished between trolling-like and non-trolling-like behavior, via criteria derived from earlier research (Hardaker 2013; cf. Appendix 1), and then we discuss the interrater agreement we

reached with our operationalization. Subsection 3.2 describes the interpretive conversation-analysis method that we applied to study trolling-like interaction.

3.1. The data

To have a valid basis for gathering the data, we adopted Hardaker's typology of six trolling strategies, as outlined above. Appendix 1 presents how we operationalized these strategies as criteria for identifying trolling-like behaviors in online discussions. Using Hardaker's descriptions of linguistic and content-related characteristics of trolling, we were able to identify observable indications of trolling-like behavior within a framework grounded in previous research.

We determined at the outset that we wanted to analyze material from political or societal and hobby- or casual-interest-based forums both, from a wide range of online sources, to increase the variety of trolling styles in our data. The online conversation spaces that we investigated were discussion areas on Reddit and comment sections of English-language online newspapers: *The Telegraph*, *The Guardian*, and *The Washington Post*. These are influential media platforms with a large readership, which renders them likely to be targeted by trolls. Our choice of platforms was designed for heterogeneity of conversation styles in the corpus, in the awareness that political ideology, context, (target) audience, popularity, accessibility, moderation practices, and language all exert an effect on the type of discussions readers are likely to engage in on a given site. Also, we focused our data collection on discussion sites that were not moderation-heavy, since service providers with strict moderation practices would have very likely deleted most posts that resemble trolling. Appendix 2 characterizes our data sources.

We acquired data from discussions that took place mainly in late 2018 to 2019. The political/societal topics we chose to look at were *climate change* and *Brexit*, and the interest-based topics were *cats* and *fitness*. Discussion of important political topics is likely to attract trolls attempting to disrupt the dialogue or incite polarized exchanges (e.g., Badawy et al., 2018; Phillips 2015). Conversations around Brexit and climate change have been especially prone to attract this type of attention (e.g., Antonio and Brulle, 2011; Humprecht 2018; Williams et al., 2015). However, apolitical and more prosaic topics too may become targets of trolling if the topic is dear to the discussion community (e.g., horses; see Hardaker 2010).

For each forum, we proceeded to read the comment sections on each subtopic, with all the individual comment *threads* (cf. Smith et al., 2000) (in all, approx. 500 threads), until we had identified at least two conversation trees on each subtopic, wherein a participant displayed trolling-like behavior. Here, 'conversation tree' refers to a branch within a thread (a thread is the full branching structure initiated by a root-level Reddit post or the whole comment section after a news article), where the branch is demarcated as consisting of the trolling-style post and all resulting branches of follow-up posts. Not all samples of trolling-like behavior were confined strictly to one tree, though: some users did not use the forum's response and

threading functions as intended, and there were some cases of the troll-like user continuing or reinitiating the discussion in another tree. Such instances were counted only once, lest such a user's behavior get over-represented in the data. We looked for activity-rich discussion trees, to uncover trolling-like conversation strategies that proved clearly able to elicit multiple responses (directly or indirectly). For the online newspapers' comment sections, this entailed approximately 8–15 response posts in the discussion tree triggered by the troll-like user, and the figure for Reddit is approximately 15–20.

Although our study of trolling-like behaviors was interpretive in nature, we found it important for the corpus to be rigorously gathered, so that it could be later extended in size via application of the same data-collection criteria. Since only the first author of this paper carried out the collection described above, we conducted interrater agreement analysis to evaluate the possibility of biased data collection. The dataset for interrater evaluation consisted of all the trees collected that contain trolling-like interaction, apart from 21 examples already seen by the other two authors, and an approximately equivalent set of non-trolling-resembling conversation trees retrieved from the same sources and topic spaces. Because the first author prepared both sets, their exact sizes remained unknown to the other two authors. The other two authors first classified 22 randomly selected conversation trees, using an early version of the coding system. Then we compared the classifications across all three authors: discussed and resolved all conflicts, and refined the coding protocol as necessary. Finally, we classified the remaining conversation trees (54 in total) in randomized order. There was 86.4% overall agreement, and the free-marginal Fleiss kappa value was 0.73, which is near the upper bound for substantial interrater agreement (range: 0.40–0.75; Fleiss 1981).

Tables 1–2 provide a more detailed description of our corpus. Table 1 presents a breakdown of the threads collected, by source and discussion topic. The counts refer to the number of (news-article-specific) comment sections where an instance of trolling-like behavior was found.

The sections from which we collected comments had approximately 1113 comments, on average, with many Brexit or climate change related sections having a

Table 1. The number of threads collected that include trolling-like behavior, by discussion topic and source (all manifested successful elicitation of reactions).

Source	Topic				Total
	Brexit	Climate change	Cats	Fitness	
reddit.com	4	7	5	4	20
telegraph.co.uk	7	6	2	5	20
washingtonpost.com	4	5	5	3	17
guardian.co.uk	2	3	3	3	11
Total	17	21	15	15	68

Table 2. The number of trolling-like strategies identified for the various discussion topics in trolling-like conversation trees within the threads collected.

Strategy*	Topic				
	Brexit	Climate change	Cats	Fitness	Total
Digress	5	11	1	4	21
(Hypo)criticize	2	1	0	2	5
Antipathize	11	25	5	6	47
Endanger	0	1	4	2	7
Shock	2	2	2	0	6
Aggress	8	7	14	12	41
Total	28	47	26	26	127

*Per the classification by Hardaker (2013)

considerably higher average comment count (approx. 1712), than sections on cats or fitness did (402).

The distribution of trolling strategies, per Hardaker's classification, manifested for each discussion topic can be seen in Table 2. The totals in Table 2 differ from those reported in Table 1 because a conversation sometimes exhibited multiple styles of trolling in parallel. Nearly all styles in Hardaker's typology (2013) were displayed in discussion of each topic selected.

Hardaker's categorization concentrates on linguistic and content-related features of trolling styles, without attempts to analyze in more detail how the various strategies are executed at the level of interaction or which actions in conversation render them effective. Therefore, once we had a corpus indexed for the trolling-like behaviors as defined by Hardaker, we proceeded to analyze the conversational interactions between troll-like users and non-troll-like users with the aid of CA.

3.2. The analysis method

We carefully combed through the final collection of 68 threads to analyze the conversation trees that manifested trolling-like behavior. This involved 1263 posts: in total, 828 from non-troll-like users and 435 from troll-like users. In applying CA concepts to our data, we were aware that CA is an interpretive research approach that requires expert human judgment. Therefore, we employed precisely defined analytical concepts and constructs in our analysis to provide concrete raw evidence for every conclusion stated.

Examining turns in interaction is fundamental to CA. For this, we utilized *adjacency pairs*. To avoid technical jargon and for ease of explanation, we will refer to adjacency pairs as action pairs or paired actions. Whatever term is used, the notion refers to a structural unit of conversation that in our case consists of a pair of posts to an online forum: a first pair part (FPP) and a second pair part (SPP). The first element

in the pair, the FPP projects and simultaneously creates expectations of a specific type of SPP. For example, a greeting-like post anticipates another greeting, a question requests an answer, and an invitation awaits acceptance or rejection (Schegloff 1968, 2007; Schegloff and Sacks, 1973). Through paired actions, participants behave in a coherent and explainable, or *accountable*, manner. Behaviors that deviate from the normative patterns recognized for action pairs typically prompt others to demand that the deviator explain the aberrant behavior – to provide an account – and clear up the confusion (e.g., Garfinkel 1967; Heritage 1984).

We chose action pairs as our main object of interest for two reasons: (1) there already is a vast body of CA research on them (e.g., Schegloff 2007), which has demonstrated their fruitfulness in analyzing how people manage actions and activities jointly, and (2) conversation analytic studies of online environments, especially studies that look at maintaining coherence, have shown that people generally are strongly oriented toward paired actions in interaction (e.g., Meredith 2019; Skovholt and Svennevig 2013). Although text-only online communication has been described as only loosely coherent when compared to face-to-face conversation (Herring 1999) and also the adjacency of a pair's two member actions on some platforms is disrupted (Giles et al., 2015; Gruber 1998; Herring 1999; Markman 2005), people still typically consider a post to have a main action that should be addressed in an expected manner (Stommel and Koole, 2010). Furthermore, prior research attests to the utility of studying conversational actions specifically in online communication (Condon and Cech 2001; Gruber 1998) – for instance, to aid in developing automated classification systems (Twitchell and Nunamaker, 2004) that identify speech acts (Austin 1962).

To annotate interactions with codes referring to specific conversational actions, we made particular use of literature on how action pairs unfold in typical conversations. Appendix 3 presents the full list of codes we used. We carefully analyzed the instances wherein social rules were not honored and discussion was disrupted. Social norms of conversation suggest that an FPP should be followed by an appropriate SPP (Heritage 1984, pp. 245–253; Levinson 1983, pp. 332–336; Schegloff 2007, pp. 13–27); that is, the FPP and SPP should be congruent with a typical action-pair type, such as question–answer (Stivers and Rossano, 2010; Stivers et al., 2010). Again, deviation could impel others to demand the deviant poster to either produce the expected response or offer an explanation (Heritage 1984; Pomerantz 1984). In the absence of an explanation, violation of the norm might also attract negative attention and follow-up questions, coupled with worries about ulterior motives (Greatbatch 1986; Romaniuk 2013). The costs for resolving norm-violation situations might be high, especially since attempts to hold the perpetrator accountable disrupt the flow of the discussion (Clark and Wilkes-Gibbs, 1986; Heritage 1984).

The notion of a proper SPP is closely connected with the CA concept of preference organization (Pomerantz 1984; Schegloff 2007, pp. 58–81), which refers

to a set of social norms articulating the preferred (and dispreferred) ways of interacting (e.g., there is a norm that accepting an invitation is preferred over rejecting it, and responding to a question is preferred over not doing so). Several preference principles may be relevant simultaneously, so participants judge which principles hold in the situation in light of prior actions. In this article, we not only analyze dispreferred SPPs but also pay attention to an even more specific phenomenon: trolling-like responses are systemically deviant – i.e., dispreferred – yet are produced as if they were unproblematic; there is no justification or hesitation. Thereby, trolling-like response patterns constitute a unique form of non-normative responding.

To analyze paired actions in interactions with troll-like users, we developed an analytical frame by synthesizing widely acknowledged conversation analytical typologies of actions (Clark and Schaefer, 1989; Stivers 2013, p. 192; Vatanen 2014). This process, informed also by our close reading of the data, yielded 13 action-pair types, which we used in preliminary analysis to annotate interactions by action-pair category. Appendix 3 presents all the action pairs found in the seemingly trolling-displaying conversation trees we analyzed, covering paired actions found in any posts, by both troll-like and non-troll-like users, and both symmetric and asymmetric pairs. We counted the pairs found in the annotated data with reference to the pair-initiating part (the FPP). Four categories proved significantly more frequent than others:

- *Question – answer*, where the FPP (a question) creates an expectation of an answer (Stivers and Robinson 2006). Other appropriate SPPs are reports of not having access to information relevant for answering and criticisms of the question's implied premises (Thompson et al. 2015).
- *Assertion – agreement or disagreement*, where the FPP makes a claim about a general state of affairs, often assuming an evaluative, personal stance to it (Vatanen 2014). Although assertions do not show strong expectations for a certain response, they often anticipate another assertion (confirming or disconfirming), consistent with the topic (Heritage and Raymond 2005; Pomerantz 1984; Vatanen 2014).
- *Accusation – admission or denial*, where the FPP points to an admission or denial as the expected SPP (Dersley and Wootton 2000, p. 387; Drew 1978). The expected response in this case is a confirmation of the action mentioned in the FPP, possibly also explicitly supplying an account for the action (Dersley and Wootton 2000).
- *Request for action – acceptance or rejection*, where the FPP normally limits the SPP to either an acceptance or a rejection (Thompson et al. 2015).

We expected to see possible trolls using the normative constraints of action pairs differently from what happens in ordinary communication aimed at common grounding. In particular, we supposed that they would reply in a condescending manner or might respond to only those questions that

furthered their ends. In the discussion of our results, we demonstrate that trolling-like responses indeed often differ from typical responses, and we consider these findings in terms of the principles of common grounding, which we found to explain the phenomenon and its success.

4. Results

Hardaker's typology of styles of trolling exhibited in online conversations not only was vital to our criteria for corpus collection but also served as a basis for the analysis itself, aimed at improving our understanding of how the use of conversational turns may differ between troll-like users and other, non-troll-like users. Our main output was identification of three ways in which troll-like participants' use of conversational actions differed from those of non-troll-like users. All three are related to what we identified as *symmetry violations* in the action pairs. *While non-troll-like users predominantly produced symmetrical turns – typically, an FPP was followed by an expected SPP – troll-like users often produced contra-normative turns: unexpected (i.e., asymmetric) SPPs.* Below, we will refer to the expected, 'proper' SPPs as symmetric and the unexpected SPPs as asymmetric.

The three asymmetric actions identified are *ignoring*, *mismatching*, and *challenging*. In ignoring, the troll-like participant does not produce a response at all – for example, leaving a question (the FPP half) unanswered. In mismatching, the troll-like user fashions an SPP that violates the Gricean maxim of relevance: misinterpreting the FPP that preceded it or producing a post without relevance in relation to the main content of the first action. Finally, in challenging, the troll-like user questions the justifications for the FPP or challenges its author.

Table 3 shows the frequencies of asymmetric actions produced by non-troll-like and troll-like users in response to FPPs expecting an answer. The symmetric entries in this table feature responses (SPPs) to FPP actions that represent completion of a recognized action pair, in some cases following requests for further information before the responder answers. The overall values suggest that trolling-like communication contains more asymmetries than non-trolling-like communication, along with fewer symmetric closures for paired actions. We did not attempt to verify this finding statistically, however, since the observations are not independent of each other: often, a conversation incorporating trolling-like behavior displayed several symmetry violations produced by the same user. We leave such comparison for later analysis and note only that troll-like users' posts appear to manifest an unexpectedly large percentage of conversational violations that disrupt grounding. In the following subsections, we will describe the three asymmetric strategies in more detail.

Table 3. Counts and percentages of asymmetric actions in our data.

Asymmetric action type	User ^a (asymmetric actions/all posts)	
	Non-trolling-like	Trolling-like
Ignoring (<i>refraining from responding to an FPP aimed at a response</i>)	15/– ^b	193/– ^b
Mismatching (<i>responding in an irrelevant manner that does not respond to the FPP as expected</i>)	31/828 (4%)	67/435 (15%)
Challenging (<i>questioning the grounds for an FPP or its author's authority or legitimacy in the discussion</i>)	82/828 (10%)	86/435 (20%)
Symmetric responses	166/828 (20%)	52/435 (12%)

^aUser type was determined in line with Hardaker's (2013) classification; see the operationalization in Appendix 1

^bNo count is possible. This represents the total (unknown) number of times a user decided not to post a comment in the conversation tree

4.1. Ignoring

Ignoring refers to deliberately refraining from responding to others' posts or significant portions thereof. This type of behavior, also reported upon as a trolling strategy in Herring et al.'s research (2002), can disrupt conversation and provoke others because it may imply cold-shouldering or create a suspicion that the interlocutor has missed or misunderstood the first action (Schegloff 2007). In our data, troll-like users baited others into participating in off-topic discussion by exploiting their desire for continuity. Ignoring others' posts or responding in only a selective manner left gaps in the conversation, which others sought to fill by highlighting the absence of replies. Troll-like users then prolonged the discussion by ignoring others' repair attempts while remaining active in the conversation in other respects. Ignoring could be seen as violating the cooperative principle, especially with regard to the Gricean maxim of quantity, in that a non-response is significantly less than what is required for informational conversation.

Our first example is an excerpt from a larger Brexit-related discussion. Here, counter-normative posting triggers angry reactions and demands for explanations from others. This diverts large amounts of space and time from the on-topic conversation. The conversation digresses into reciprocal flaming and repetitive posting in reaction to B's trolling-like behavior. In example 1, the discussion is initially disrupted by B's unexpected and provocative flaming (post 2). In post 3, the troll-like user (B, whose content is in boldface) is asked for justification for the unexpected aggression and to explain appropriating the nickname of another forum member (referred to as E in the example; see posts 8 and 10). However, B ignores these pleas while remaining active in the conversation in other ways (post 6).

Example 1: Use of the ignoring strategy in a conversation involving aggressive trolling-like behavior, in discussion of *The Telegraph*'s article 'Theresa May told "you are the problem" by backbenchers furious over Brexit paralysis as they urge her to go for good of the party,' from 9 April 2019.

Post	User	Post content
1	A	The nation state is past its sell by date.
2	B	@A You're really exposing yourself now Moshe.
3	C	@B @A Why call him that? <i>[1 side-comment omitted]</i>
4	D	@C Because C, this B is pretending to be another B <i>[forum member E]</i> . The latter has a brain. The former is an anti-semite. <i>[1 side-comment omitted]</i>
5	F	@D @C E is a white-hater. B is not.
6	B	@D @C Antisemite? How did you arrive at that? <i>[2 posts omitted: one restating that B is not E, the second being a side-comment]</i>
7	B	@D @F How can a person be "anti-group of languages?" what a bizarre accusation.
8	D	@B Your comments display that and even C has noticed. Why did you copy E's name? It doesn't confer her intelligence to you. <i>[1 side-comment omitted: debate about what constitutes anti-Semitism]</i>
9	B	@D @B Display what? Back up your accusation. <i>[2 posts omitted: debate about what constitutes anti-Semitism]</i>
10	D	@B I don't need to as you have already done so in your comment history. Now tell me, why did you copy E's name? You're still thick as s'hit.
11	C	@B @C @F @D Calling him "Moshe" as a generic Jewish name is appalling. Please don't go in that direction.

Viewing the flame (post 2) and the alleged nickname theft as things they needed to resolve so that the original conversation could resume, other participants continue speculating on the motives behind B's actions and insist on holding B accountable for the behavior displayed. They attempt this by addressing B directly, referring to B's selection of words, and describing past actions. Two distinct kinds of normative expectations are relevant here: one should (1) provide a response to an FPP directed at one and (2) give an account when identified as accountable. Contrary to the norm, B does not engage in resolving the breakdown; instead, B works against its repair by evading accountability, directing attention elsewhere (post 6), and later producing further provocation (e.g., post 7). The provocation displays that B has not left the discussion and *should* be engaged and ready to account for the actions (e.g., see Antaki et al., 2005). This, in turn, suggests that B is ignoring some posts deliberately.

Although B systematically acts in an incoherent manner and the others' need for explanations is apparent, other users also actively contribute to eroding the quality of the interaction (e.g., with posts 4 and 8). Their requests or demands for explanation are neither very constructive nor polite (see post 10). Thus they 'take the bait' and co-create trolling with B. They could simply have pointed out that the comment was inappropriate, then left B alone after observing B's lack of willingness to cooperate;

instead, their aggressive retaliation breeds further disruption. This too undercuts common grounding. Also, B's choice of nickname acts as a trigger for interaction, whether deliberate or not. Such nickname choice is another possible trigger for collaborative creation of trolling, if others interpret it as deliberate provocation and take the bait (on nicknames, see Lindholm 2013).

In example 2, a different user B¹ from the poster in example 1 successfully aggravates participants in a mainly peaceful pet-related discussion. The common juxtaposition of cat vs. dog is already inscribed in the title of the original article, and, by utilizing a message comparing the two, B creates an antagonistic context in which owning a cat would imply a man's homosexuality.

Example 2: Ignoring accusations of improper behavior in a conversation that includes aggressive/antipathetic trolling-like behavior, in discussion of The Washington Post's 'Dog owners are much happier than cat owners, survey finds,' from 5 April 2019.

Post	User	Post content
1	A	I have two cats who I love. My girlfriend has a dog who I love. So I guess it's ok to love them both. What's especially nice is that they all seem to love me right back.
2	B → A	Ugh. A man with cats is the worst! Just come out of the closet already.
3	A → B	I thought your earlier post was kind of obnoxious. Now I know exactly why I thought that. Thanks for clearing things up.
4	B	LOL! Cat people are so easily offended....
5	A → B	Ever ask yourself why you go out of your way to offend people? Well, wait – maybe it's not out of your way at all.
6	B → A	It's Friday. Work is slow. They made the mistake of giving me internet access....
7	C → B	They made the mistake of giving you a job.
8	B	LOL! So many offended cat people. I LOVE THIS!!!!!!
9	D	Desperate for attention and affirmation. Get yourself a drooling dog. [2 posts omitted: comments similar to post 9, criticizing A's behavior]
10	E	Spot-on. He's a miserable, immature attention-seeker who needs to insult innocent people and animals to feel validated. [2 posts omitted: comments similar to post 10]
11	E → B	Dog person here: you're obnoxious and need to examine why you delight in insulting people who seem perfectly nice. And stop insulting gays, too.
12	F → B	(Edited) Why are you being so ignorant? I have not read one decent comment from you. By the way, I am a dog person.
13	G → B	But, when he comes out, would he be a lesbian?
14	B	Hopefully, for his girlfriend's sake.

¹ The users in all our examples are designated alphabetically, starting with A.

B systematically ignores accusations of contra-normative behavior: posts 3, 13, and 14 are directed at B and call for a response. However, these posts are either completely ignored (13 and 14) or countered by implying that their authors are overly sensitive. The audience takes B's bait and collaborates in authoring offensive and contentious posts that further contribute to a negative atmosphere. Consequently, the discussion digresses into flaming and acts of retribution, while B keeps hampering any possible conclusion of the conflict, by feeding the argument with additional asymmetric posts and provocation (e.g., post 14). Thus, poster B also displays active selectivity in choosing which posts to respond to. Choosing to ignore some posts and B's manner of responding in, for example, post 14 provide evidence of systematic violation of coherence and accountability norms: instead of responding to questions such as that in post 12, B opts to respond to post 13, which continues to ridicule A and supports the trolling. In essence, B manages to prolong the conflict by working against common grounding. Rather than try to resolve problems by producing an expected reply (acknowledging the actions' impropriety and/or giving an account), B evades criticism for the antisocial behavior. Hence, the conflict does not get resolved.

In the above examples, we can see that the ignoring strategy is highly effective in protracting fruitless exchanges by luring others into requesting explanations and proper answers that could make sense of the troll-like user's ambiguous behavior. While others dutifully respond to the troll-like user's posts, the disruptive user replies only to selected messages. Leaving posts unanswered creates information gaps that disrupt common grounding. Also, the reciprocity of a discussion is compromised when troll-like users refuse to provide explanations for their offensive actions.

4.2. Mismatching

In the second strategy that we identified – mismatching – the troll-like user either misinterprets/misrepresents the FPP or responds in a manner that is irrelevant with regard to the central point made in the FPP half. Asymmetric responses of this nature are contra-normative (Heritage 1984; Levinson 1983; Schegloff 2007), and they lead to breakdowns of conversation by disrupting the coherence of meaning-making. In particular, they violate the Gricean maxim of relevance: the first action anticipates or makes relevant a restricted set of actions, to which the irrelevant and often confusing response is not a fitting or understandable match.

For example, the excerpt below comes from a conversation revolving around cats and their contributions to overcoming social anxiety. In post 3, A displays trolling-like behavior by misinterpreting post 2.

Example 3: Use of the strategy of mismatching in a conversation involving aggressive trolling-like behavior in discussion of *The Guardian's* 'Experience: I'm a full-time cat sitter,' from 24 May 2019.

Post	User	Post content
1	A	There are two kinds of people in this world. People who live with cats and people whose houses don't reek of cat piss.
2	B → A	Such a stunning insight. Did you come up with that all by yourself?
3	A → B	The first part of the aphorism is quite common. The second part is an observation that a lot of people whose houses don't reek of cat piss tend to experience. So the answer to your question is yes and no.
4	C → B	"Such a stunning insight. Did you come up with that all by yourself?" No, of course not. it's an old internet trope.
5	D → A	Lots of cats will go outside to do their business. They are actually very good at sticking to the same area outside i.e. they don't like to randomly poop all over the place. Some people will put a litter tray outside. But I agree, if you have an indoor cat it's likely there will be at least some cat pee smell as the ammonia is so strong.
6	E → A	But my cats prefer to shit and piss in your garden so I don't know how this can be true.
7	D → D	* an indoor cat that uses an indoor litter tray
8	F → A	I have a cat. My house doesn't reek of cat piss. Perhaps it's the company you keep.

The expression in post 2 'did you come up with that all by yourself?' is commonly used to point out an improper or foolish statement. Here, a symmetrical SPP might, for example, have provided an account of A's reasoning process; however, A displays misinterpretation of post 2 as an information-seeking question. Thus, A resists the demand to answer the accusation and deflects any criticism of the problematic behavior. This has an impact on other participants' responses: though some post sincere replies to A in order to work toward common grounding (e.g., post 5), most are aggravated into accusing A of disagreeable behavior (post 2) or posting insults (posts 6 and 8). If the audience had responded largely neutrally, in the manner of post 5, or had casually ignored A's provocative comment, the posters might have succeeded in alleviating the strain. In contrast, content such as posts 2 and 8 only exacerbates the trolling or continues such behavior instead of mending the fractured common ground. In this way, by baiting others to reciprocate, A renders the conversation a regressive series of finger-pointing and name-calling actions. Later in the discussion, A's further provocation and lack of accountability protract the others' futile attempts at grounding all the more, so closure cannot be achieved.

In our data, participants exhibiting trolling-like behavior often utilize misinterpretation to dodge a key observation made in an earlier post. A case in point is example 4, below, wherein B acts in a trolling-like manner by derailing a discussion of the effects of Brexit on the British economy. Its transformation into an aggressive argument over racism begins with this comment by B: 'It's almost like letting a bunch of bitter racist white losers vote for shit ruins your country.'

Example 4: The mismatching strategy used in a conversation containing digressive trolling-like behavior, under ‘UK economy shrinks by four times as much as predicted as Brexit paralysis takes hold,’ in Reddit’s r/worldnews, in June 2019.

Post	User	Post content
1	A	You: nope I’m not someone how calls everyone racist. Also you: the only reason the two biggest political movements in recent times happened is because RACISTS!!!!
2	B	You don’t have good analytical ability apparently.
3	A	Did you or did you not just boil the 2016 election and the brexit vote down to racism? Am I missing something? Because that’s exactly what you did
4	B	Yea, it’s white nationalist bullshit

The second post by B (post 4) demonstrates mismatching by misinterpretation: while post 3 is essentially an accusatory-question FPP, post 4 is an asymmetric response, better matching an information-seeking question. This intimates that B either does not understand post 3 or is willfully ignoring the point made in it. In either case, post 4 is not an expected response to post 3. Post 4’s inadequacy as a response is evident from A’s later actions (data not shown): re-articulating the accusation, to demand an admission or denial. This is done with greater aggression, manifesting frustration with post 4. Still, B resists the demands for accountability, refusing to cooperate and pursuing another agenda. At the same time, the contributions by A are quite face-threatening, calling out B’s behavior and intentions in very direct and offensive terms. Therefore, A’s reactions do not seem to offer a constructive footing for exchanging ideas and achieving common grounding. Arguably, A therefore contributes to the conversation’s breakdown. Again, sometimes a poster may not have intended to troll but trolling behavior gets triggered by the situation, thanks to others’ overly eager or aggressive reactions.

Mismatching was manifested in another way also, ridiculing the first action by producing responses utterly unrelated to the tone of the first post or its main point. In example 5, participants in a discussion in Reddit r/worldnews have been debating the economic effects of climate change. Behaving in a trolling-like manner in post 1, A leads the conversation into disarray by sniping at other users and the topic, then refusing to behave civilly and acknowledge others’ points properly.

4.3. Challenging

The third asymmetric response strategy we identified involves challenging the justification behind the initial actions. Challenges can be found in (normal) spoken conversation too (e.g., Heritage 2012; Thompson et al. 2015), so it was unsurprising that our dataset features a few challenges in non-trolling-style turns. Every challenge carries possibly face-threatening implications (cf. Brown and Levinson 1987; Goffman 1967, p. 37); however, troll-like participants used this tool in a more systematic and provocative manner. Their challenges often breached the Gricean maxim of manner. We identified two types of challenging response: attacking the grounds for an earlier post and attacking another user's authority. We discuss both below.

4.3.1. *Challenging the grounds for a post*

Challenging the grounds for an action often involved countering one question with another, thus casting the first post as unnecessary or invalid while placing the challenger in an authoritative position (see also the example of question-trolling described by Zvereva 2020, pp. 114–116). This was done also by means of such assessments or assertions as ‘What a dumb post’ or by redirecting the FPP back at the original poster (e.g., A: ‘You lack understanding of what you’re talking about.’ – B: ‘No u’). In example 6, B challenges A’s post by meeting a question with another question.

Example 6: Challenging by questioning the grounds for an earlier message. An excerpt from a conversation containing digressive trolling-like behavior in discussion of *The Washington Post*’s ‘Adrift in the Arctic,’ from June 2019.

Post	User	Post content
1	A	B: Are you a scientist? Thanks.
		<i>[3 posts omitted: responses to other posts, unrelated to post 1]</i>
2	B	(Edited) Dear A, Does it matter. If so how so? What I know about science is it is extremely equal to everyone of us. It doesn't matter who you are, it doesn't matter where you from, it doesn't matter what kind of diploma you have. The only thing that matters in science is scientific evidence.

Earlier in this conversation, other users provided evidence and counter-arguments to refute B’s provocative assertions regarding climate change. Several times, B appealed to the superiority of scientific evidence/argumentation for meaning-making within the debate, mostly to back B’s own claims including information that counters information commonly presented in the thread as accepted. In post 1, Poster A asks whether B has sufficient educational background to refute the evidence provided. With post 2, B does not produce a symmetrical SPP (i.e., one with any information on B’s scientific background), instead challenging the question’s grounds with a counter-question to render it irrelevant. The questioner is in a position

to place constraints on what the next person should do (Sacks 1992, p. 54; Stivers and Hayashi, 2010); by directing the question back to the first person, B takes control of the interaction here. After B's challenge, a futile battle for authority erupts between B and other users (not shown in the example), who engage in a tangential debate over B's educational background. The reactions of the audience are noteworthy in other respects too. For instance, A's post 1 is somewhat confrontational. The anticipatory 'Thanks' after the question may frame it more as a rhetorical question or challenge than a sincere question, or the poster's thanks may be taken as sarcastic. Although many participants in this conversation (beyond example 6) attempt to engage fairly neutrally in conversation with B, many posters, A among them, can be seen as posting provocative responses, possibly rendering the discussion even more negative in tone.

4.3.2. *Challenging someone's authority*

The second way of issuing a challenge is to call another user's authority into question, either by questioning that user's epistemic authority on a topic or by questioning the poster's legitimacy as a valid or sincere participant in the discussion. In example 7, a discussion originally revolving around cats, B challenges another user via accusations of trolling – i.e., questioning that user's sincerity within the discussion.

Example 7: Challenging by questioning another user's authority, as displayed in a discussion tree involving aggressive trolling-like behavior, found under 'Love our baby girl more than anything. I've never been able to understand why they're considered bad luck,' in Reddit's r/cats, from April 2019.

Post	User	Post content
1	A	<i>cough</i> Still waiting on them sources...
2	B	The troll is interfering.
3	A	Agreed! So stop it already and cite your sources?
4	B	I already stated it troll.

Here, A keeps insisting that B's assertions should be proven with proper references, using an indirect request-for-action FPP in post 1. In response, B resists this request, by accusing A of being a troll and, rather than providing the requested information, altogether denying A's authority to ask for references. The situation escalates into reciprocal accusations of trolling, by means of which B is able to divert attention from unwanted questions. Such retaliatory accusations of trolling are quite common in online disputes (e.g., Knustad 2020), and they can be seen as another mechanism of co-creating trolling. In some cases, such as this one, the accusations themselves are a form of trolling; the user acting in a trolling-like way levels them against a sincere poster.

In our data, questioning someone's legitimacy as a valid participant in the discussion proved to be a highly provocative means of challenging that person, especially in cases in the 'aggress' category (see Appendix 1). Overall, the challenges created an atmosphere of uncertainty in the discussion space. By exploiting challenges to someone's authority, troll-like users incited quarrels over authority and prolonged regressive exchanges. Besides violating the Gricean maxim of manner, this fundamentally obscured honoring of the maxim of quality in the discussion space – it cast doubt over participants' sincerity and their (epistemic) authority on a given topic. This muddling of the apparent motivations and identities of several users in the discussion space frequently resulted in bystanders not being able to distinguish sincere contributors from deceptive troublemakers. Moreover, the time and effort costs for reaching common grounding to resolve these breakdowns were often high.

5. Discussion

The core contribution of the study is our conversation analysis treatment of the turn-taking strategies used by troll-like participants in online discussions. We identified patterns used to degrade online conversation and found that troll-like users succeed in disrupting conversations by hindering common grounding – the joint search for conversational closure and maintenance of mutual understanding.

When closely examining the action pairs in the corpus of online conversations capturing observable characteristics of trolling-like behavior per Hardaker's typology, we found that, across all political/societal and leisure-related discussions, troll-like users disrupt common grounding by deviating from expected conversational norms. Instead of continuing conversations in a way that would afford symmetry between an action pair's first part (the FPP) and the second half (the SPP), they create asymmetries. Through our conversation-analysis approach, we uncovered several conversation strategies that can highly effectively derail conversations, frustrating common grounding or a satisfying closure. Of these techniques (characterized above as ignoring other users' posts that anticipate a response, posting various mismatching responses, and challenging other users' legitimacy or the grounds for their comments instead of addressing their posts' content), only the first – ignoring – has been identified by scholars, in a brief observation by Herring et al. (2002). Certainly, none of them has been analyzed in detail.

The interaction patterns we pinpointed also, importantly, illustrate the collaborative nature of trolling. By not following the rules of conversational symmetry, troll-like users bait others to respond. While one user might perform counter-normatively in a relatively systematic manner, thereby

disrupting the coherence of the conversation, trolling is a joint creation (as noted also by Cook et al, 2019), and anyone may end up contributing to it (cf. Cheng et al. 2017). Thus, trolling as a phenomenon emerges from community experience and culture: both for perceiving/identifying trolling behavior in the given context (cf. Sanfilippo et al. 2018) and for contributing to the trolling event or prolonging it, collaborative effort is necessary. As a good troll knows, ‘it takes two to tango.’

The maxims for what is expected in conversation frame our results well. We found that they often do not hold in conversations in which troll-like participants take part. By anchoring our findings theoretically in light of the Gricean cooperative principle, we were able to conclude that trolling-like behavior breaches the maxims of quantity (in ignoring), relevance (in mismatching), and manner (in challenging). Finally, when the trolling-like behavior is seen as an attempt to deceive other users, all of these strategies flout the maxim of quality, the notion that what is said should be true. Overall, the troll-like participants’ actions in the discussions display a strategically uncooperative orientation.

5.1. Limitations of the work

Our analysis was focused on only a subset of trolling-like phenomena. We relied on Hardaker’s definitions in conceptualizing six distinct trolling strategies to guide our collection of data. Hardaker’s framework is not comprehensive, so identifying trolling by means of it may have blinded us to other forms of trolling. In addition, although we sought heterogeneity by analyzing both political and leisure-interest-based discussions and by gathering material from four distinct online forums, collecting data from only these sources may have exposed us to a limited portion of the full set of trolling-like strategies possible. Also, while Hardaker’s categorization is useful for studying differences in trolling-like behaviors, its operationalization requires interpreting the categories in new contexts, quite different from those she cited herself. Further work could, accordingly, lead to the expansion of some categories or changes in the interpretation of their precise meanings. However, in light of our substantial interrater agreement, we concluded that we succeeded well in operationalizing the characteristics of trolling-like behaviors with the coding system in Appendix 1.

We examined the data only through the lens of conversation analysis, with special regard to action pairs and their symmetry violations as our main concepts. While addressing the research gap, this did restrict the possibilities for noticing any trolling mechanisms that cannot be identified via the lens of action pairs. Analysis of other action types or alternative specific CA phenomena might yield further findings. For instance, studies could concentrate on ‘third positions,’ turns that respond to the SPP.

Another possible factor is that preference principles in online conversations may differ somewhat on the basis of context, just as they diverge from those in face-to-face interactions in several ways, since the media offer very different affordances. For example, no one is held accountable for not responding to a post directed at a forum group as a whole. Furthermore, forum participants may vary greatly in sociocultural background, and research attests that the weight accorded to specific preference principles depends partly on cultural group (e.g., Goodwin and Goodwin, 1987).

Finally, while many of these limitations imply that other trolling-like conversation strategies have yet to be identified, the ones we found already provide ample evidence of effective disruption. Such strategies' effectiveness can be explained in theoretical terms by drawing from psycholinguistic research on common grounding and on ethnomethodological and conversation analytic understanding of normative human behavior.

5.2. Conversation analytic approaches to trolling

One considerable benefit of our theoretical and methodological approach is its agnosticism to troll-like users' true identity and intentions. Thus far, research into trolling has been preoccupied with studying it as intentional behavior (see Hardaker 2010). Therefore, the problem of creating a corpus that represents true trolling has been inescapable. After all, the identity of a suspected troll is seldom known; neither can the intentions behind even an avowed troll's behavior be ascertained with any certainty. For scholars analyzing online material, it has proven highly challenging to judge whether a given user was trying to troll others or not. Since trolling is a deception game (Donath 1999) wherein the troll's success relies largely on an ability to feign sincerity, extensive datasets wherein the trolls are conclusively identified are extremely difficult to generate.

By putting relatively little stress on the intentions behind the turn-by-turn interactions (Hopper 2005), conversation analysis is not hampered by such problems. One can pinpoint trolling-like behaviors in data by analyzing the effects of conversational turns on subsequent turns, especially with respect to repairs and attempts to maintain common grounding. For example, the effects of misinterpreting a post are visible as the discussion unfolds, and misinterpretation can be viewed as deliberate if the suspected troll does not self-correct or react in another normatively expected manner once other users have pointed out the improper behavior. In this light, behavior patterns can be identified and analyzed independently of whether the user intended to troll others. This creates opportunities for considering trolling-like behavior through its observable behavioral characteristics instead of its mind-internal, intention-bound nature. Thereby, trolling-like behavior is rendered much more amenable to analysis, especially since research traditions in

conversation analysis offer a strong methodological and theoretical starting point for such work. Particularly important is the EM principle of a moral obligation of being accountable and understandable to others in one's interactions (Garfinkel 1967; Heritage 1984). Trolling-like behaviors can be seen as violations of this fundamental norm. This explains their great harmfulness in online conversations.

We see potential for research to develop a new operational definition for trolling, or trolling-like behavior, one that does not lead to problems similar to those accompanying the prevailing approach – especially the requirement for theorizing about trolls' identity and intentions, neither of which a researcher can reasonably ascertain. This work should be informed by studies with larger and more comprehensive corpora, to test whether our findings withstand statistical analysis.

5.3. Implications of the findings – positive and negative potential

Online social media are a catalyst for social unrest, expressions of misogyny, and other negative behavior. With regard to solving these problems, our findings might have negative implications, but they might also point to ways forward. We have shown that, at least to some extent, trolling is an activity that can be traced, recognized, and labeled. This deeper understanding of the specific techniques that lure others into meaningless fights may be put to harmful purposes, but it can also assist in detecting and preventing disruptive behaviors.

There are two possible negative outcomes of work such as ours. Firstly, for trolls, our research might point to more effective ways to disrupt online conversations. A more far-reaching possible outcome is that research of this sort may identify means by which automated trolling mechanisms could be created and honed. If trolling-like behaviors can be codified as repeatable conversation patterns, along the lines of those developed for ordinary chatbot design (see Moore and Arar, 2019), chatbots could emerge that take part in conversations as seemingly competent partners. Their trolling may be sophisticated enough to pass as genuine, sincere behavior. Therefore, it may be impossible to straightforwardly ignore the trolling attempts.

We find this scenario plausible, since trolling does not necessarily require the levels of sophistication or interpersonal sensitivity that mediated social interaction generally does. In essence, the effective trolling strategies that we have examined are a set of antisocial patterns violating the Gricean cooperative principle and are similar to the 'breaching experiments' discussed by EM scholars (Garfinkel 1967). Such patterns could lend themselves well to automation. So far, it has been fairly easy for people to recognize when they are interacting with a chatbot; likewise, chatbots that troll – trollbots – have seemed easy to ignore. However, were it possible to develop more capable trollbots that pass well enough as human, we could imagine a dystopian future wherein low-cost trollbots are planted in online forums to systematically disrupt and manipulate civil discussion.

On the positive side, awareness of trolling mechanisms can feed in to work on reducing the ripples from trolling. Arguably, trolls themselves and other malicious actors are already aware of the types of conversation strategies that derail discussion. For countermeasures, it may be possible to educate users of some discussion forums about conversation strategies that often disrupt successful exchange of ideas. With greater user awareness of such strategies, possible trolling attempts might be more readily detected and defused. However, efforts at thus countering trolling-like or other disruptive behaviors have not proven sufficient in the nearly 20 years since Herring et al. put the idea forward (2002, p. 381). It would seem, then, that automated moderation will eclipse them in attempts to keep online conversation civil.

Automated moderation to address trolling-like behaviors seems feasible. If, as we speculate above, trolling can be automated, it should be possible to recognize it automatically too. We argue that identifying measurable systematic patterns in interaction and conversational coherence – such as the asymmetric responses described in this paper – should offer a means for developing automatic identification and mediation of trolling and other disruptive behavior. Various moderation approaches are possible. For example, posts that manifest characteristics of known trolling strategies, such as asymmetric responses, could be flagged as possible attempts at disrupting the conversation or trolling. This flag, such as a warning symbol, could be supplemented with an explanation of grounds for thinking the post might be written by a troll and of what harm it could produce in the subsequent discussion. Others taking part in the conversation may thus be primed to look out for such attempts. An automatic moderation system could also provide users with warnings or suggestions as they write posts, in case they start composing a message in an offensive tone or without responding to another user's question or comment. Naturally, one can imagine far more direct methods of moderation, including simply blocking individual posts or outright banning a user whose posts recurrently display characteristics of trolling.

So far, research has not developed effective methods to prevent online trolling. At present, both academic and practical research efforts seem to be lagging far behind the methods and tactics of harm-bringing participants in online discussion. It is our sincerest hope that the findings presented in this paper lead to more effective means of detecting and preventing trolling attempts in everyday online interaction, to eliminate their vast potential for harm.

Acknowledgments

This work has been supported by Academy of Finland (grant nr. 320694).

Funding

Open access funding provided by Aalto University.

1. Appendix 1

The coding system for distinguishing between trolling-like and non-trolling-like behavior.

We used the following principles to gather a corpus of trolling-like behavior from online discussion forums.

Requirements for coders: The annotator must be familiar with online trolling as a phenomenon and the relevant literature (at least Donath 1999; Fichman and Sanfilippo, 2016; Hardaker 2010; Hardaker 2013; Herring et al., 2002; Phillips 2015).

Exclusion criteria: Participation in online conversation is classified as trolling-like behavior if its characteristics match the operationalization of Hardaker's trolling strategies presented below (see the table) – except for the following cases, for reason of protecting the classification from being too inclusive and thereby leading to a corpus that contains some examples of behaviors that are not trolling-like:

1. The possibly trolling-like post provokes no responses.
2. The trolling-like behavior does not elicit a reaction demonstrating that the behavior was somehow problematic to other participants. The case might involve, for example, an unsuccessful trolling attempt or behavior that is not trolling-like at all. Paying attention to other users' reactions helps one exclude cases of, for instance, prompting irony or exaggeration that actually manifests a commonly agreed type of humor or good-natured teasing rather than trolling.
3. Repeated posts by the same participant provoke no responses. Even if the participant makes several posts that could be deemed trolling-like, these are not considered trolling-like within this study if they fail to elicit responses (per rule 1).
4. A single possibly trolling-like post lacks trolling-like continuation. If a participant (e.g., in a heated discussion) posts one message that could be considered trolling-like but does not follow up with additional messages of that type, the post is not deemed trolling-like, even if provoking a response.
5. Disinformation is posted. Distributing fake news in a forum is not in itself trolling-like behavior in our classification. It must also involve some characteristics of categories in the table. That is because it represents online trolling of a sort other than what Hardaker's categorization addresses.
6. The discussion involves disagreeing and angry exchange of messages. Only if the exchange also involves characteristics of the six categories presented below (e.g., aggression such as name-calling or antipathy such as unsupported challenging of another party's motivations) are angry messages considered to show trolling-like behavior.

Inclusion criteria: Trolling-like behavior is recognized and classified in line with Hardaker's (2013) classification as presented below. In many cases, a trolling-like behavior may display characteristics of several categories. For instance, aggressing might appear in combination with antipathy trolling, and hypocritical remarks on punctuation might also be examples of digression.

	Trolling-like behavior	Not trolling-like behavior
Digress	<ul style="list-style-type: none"> - Posts that lead others away from the original topic of discussion, often into pointless, frustrating, or circular discussion that does not advance the discussion much (see Hardaker, 2013), as indicated by negative reactions from others. - Multiple posts that 'spam' the forum with nonsense, repeated content, or posts that initiate or prolong cascades (e.g., reciting numbers or the alphabet or replying to other posts with rhymes). 	<ul style="list-style-type: none"> - The poster returning to the original discussion or allowing others to discuss the actual topic.
(Hypo)criticize	<ul style="list-style-type: none"> - A poster criticizing others because of spelling, grammar, punctuation, or other irrelevant details. - Cases of the critical poster actually making mistakes of the same sort. 	<ul style="list-style-type: none"> - A poster noticing another user's (e.g., punctuation) error and correcting it but not in a provocative manner – e.g. no aggressive word use or name-calling is involved, etc.
Antipathize	<ul style="list-style-type: none"> - A poster proactively introducing a new sensitive/antagonistic context that may not have been the original topic of discussion. - A post building on an opportunity for triggering arguments in the earlier conversation and creating a context for further conversation that exploits people's sensitive spots. 	<ul style="list-style-type: none"> - A poster introducing a controversial point in a discussion (e.g., whataboutisms) but providing reasonable evidence or arguments to back the claim. The poster may take part in the debate that follows.
Endanger	<ul style="list-style-type: none"> - A poster disseminating bad advice or offering a dangerous example for others to follow while giving an impression of being ignorant of its harmfulness. - Despite others' pleas, continuing to give bad advice that is not accepted. 	<ul style="list-style-type: none"> - A poster offering bad advice but correcting this when others point out the problem, after which others accept the correction.

(continued on next page)

(continued)

	Trolling-like behavior	Not trolling-like behavior
Shock	<ul style="list-style-type: none"> - Posts about a sensitive or taboo topic, such as religion, death, human-rights abuses, animal rights, or politics. - Posts that are inappropriately thoughtless or hurtful in a sensitive, upsetting, or emotion-fraught situation. - A poster exploiting an existing sensitive context or normative frame to trigger arguments or shock people. 	<ul style="list-style-type: none"> - If evidence or an argument is given, even content that might seem quite shocking to some – this is not necessarily trolling.
Aggress	<ul style="list-style-type: none"> - A poster openly provoking others into retaliating – e.g., using ad hominem arguments to start a fight. - A poster commenting aggressively on the topic (which may be a person or group of persons, such as a politician or a minority group), prompting others to defend the target. 	<ul style="list-style-type: none"> - Aggression that is sufficiently justifiable in light of earlier events, such as another person’s insulting posts. - Accidental insults (the person responsible apologizes).

2. Appendix 2

Data sources used in the project.

Source	Description
reddit.com	Average users are predominantly male and USA-based, and the average age is 18–29 (Pew Research Center, 2016). Content is characterized as often discriminatory and violent (Hankes, 2015; Lewis, 2015), but there is great variation between ‘subreddits’ and in their norms (Chandrasekharan et al., 2018; Fichman and Sanfilippo, 2016, p. 150), which tends to guide conversations. While discussions in r/worldnews and r/ukpolitics are quite critical and often ironic or sarcastic, with occasional flaming, those in r/cats and r/fitness, where discussion revolves around the hobby, are less so. Conversation trees’ structure in combination with the peer upvote/downvote (karma) system, albeit highly subjective, makes it harder to dominate or derail a conversation completely.
telegraph.co.uk	Users are mainly UK-based, and the average age is approximately 35–39 (Telegraph, 2012). In general, conversation can sometimes be quite civil although critical, but some aggressive behavior and provocative content appear quite often. Unless they breach community guidelines, all posts in the comment section are shown in order, with responses to a specific post indented and marked with @[OP’s name]. The structure of the comment sections makes any polarizing, provocative, or digressive content visible to all and enables disrupting discussions and hogging space via circular arguments.

(continued on next page)

(continued)

Source	Description
washingtonpost.com	The user base is mainly USA-based (Washington Post, 2007), and ages are likely to be higher than Reddit users'. In general, conversation tends to be polarizing and aggressive, especially with regard to political news items. The comment section is similar in structure to <i>The Telegraph's</i> except that it lacks the OP's name.
guardian.co.uk	An average user is typically UK-based, but <i>The Guardian</i> has a vast global audience. At roughly 37–38, the average user age is higher than Reddit's (Guardian, 2011, 2013). Conversation tends to stay civil fairly often, but some aggressive behavior and provocative content are seen occasionally. The structure is similar to <i>The Telegraph's</i> , and responses are marked with an arrow from the poster's nickname to the OP's name.

3. Appendix 3

All action-pair categories in the data.

Action pair	Trolling style						Total ^a
	Digress	(Hypo)-criticize	Anti-pathize	Endanger	Shock	Aggress	
<i>Major categories</i>							
Assertion–assertion	170	16	208	28	3	298	545
Accusation–admission/ denial	126	6	107	16	4	206	317
Request for information – answer	85	10	76	17	2	114	221
Request for action – acceptance/rejection	40	2	33	15	1	67	102
<i>Minor categories</i>							
Proposal–acceptance/ rejection	2	0	7	7	4	3	15
Compliment–agreement/ disagreement	0	0	0	4	0	3	6
Greeting–greeting	0	0	1	1	0	1	3
Farewell–farewell	0	0	2	1	0	0	3
Offer–acceptance/ rejection	0	0	1	1	0	0	2
Apology–acceptance/ rejection	0	0	1	0	0	2	2
Invitation–acceptance/ declination	0	0	1	1	0	0	2

(continued on next page)

(continued)

Action pair	Trolling style						Total ^a
	Digress	(Hypo)-criticize	Anti-pathize	Endanger	Shock	Aggress	
Thanks–acceptance/ rejection	1	0	1	1	0	1	1
Summons–answer	0	0	0	0	0	0	0
Total	424	34	438	92	14	695	1219

^aThe table shows the total number of pair-initiating parts (FPPs) in our data that initiate an action pair. The pairs are presented by pair type and were initiated by both troll-like and non-troll participants. The number of paired actions was counted per FPP. Some of these involve two trolling styles; since the totals in the last column do not count these cases twice, they are not direct totals of all instances of the various trolling styles

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akhtar, Shazia; and Catriona M. Morrison (2019). The prevalence and impact of online trolling of UK members of parliament. *Computers in Human Behavior*, vol. 99, pp. 322–327. <https://doi.org/10.1016/j.chb.2019.05.015>
- Antaki, Charles; Elisenda Ardévol; Francesc Núñez; and Agnès Vayreda (2005). “for she who knows who she is:” Managing accountability in online forum messages. *Journal of Computer-Mediated Communication*, vol. 11, no. 1, pp. 114–132. <https://doi.org/10.1111/j.1083-6101.2006.tb00306.x>
- Antonio, Robert J.; and Robert J. Brulle (2011). The unbearable lightness of politics: Climate change denial and political polarization. *The Sociological Quarterly*, vol. 52, no. 2, pp. 195–202. <https://doi.org/10.1111/j.1533-8525.2011.01199.x>
- Austin, John L. (1962). *How to Do Things with Words*. Oxford, UK: Oxford University Press.
- Badawy, Adam; Emilio Ferrara; and Kristina Lerman (2018). *Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign*. ASONAM'18: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, New York, NY: IEEE, pp. 258–265. <https://doi.org/10.1109/ASONAM.2018.8508646>

- Baron, Naomi S. (2000). *Alphabet to Email: How Written English Evolved and Where it's Heading*. New York, NY: Routledge.
- Bennett, W. Lance; and Steven Livingston (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, vol. 33, no. 2, pp. 122–139. <https://doi.org/10.1177/0267323118760317>
- Bishop, Jonathan (2014). Representations of ‘trolls’ in mass media communication: A review of media-texts and moral panics relating to ‘Internet trolling’. *International Journal of Web Based Communities*, vol. 10, no. 1, pp. 7–24. <https://doi.org/10.1504/IJWBC.2014.058384>
- Brown, Penelope; and Stephen Levinson (1987). *Politeness. Some Universals in Language. Studies in Interactional Sociolinguistics 4*. Cambridge: Cambridge University Press.
- Cech, Claude G.; and Sherri L. Condon (2004). Temporal properties of turn-taking and turn-packaging in synchronous computer-mediated communication. *HICSS'04: Proceedings of the 37th Annual Hawaii International Conference on System Sciences*. New York, NY: IEEE, pp. 1–10. <https://doi.org/10.1109/HICSS.2004.1265282>
- Chandrasekharan, Eshwar; Mattia Samory; Shagun Jhaver; Hunter Charvat; Amy Bruckman; Cliff Lampe; Jacob Eisenstein; and Eric Gilbert (2018). The Internet’s hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *CSCW'18: Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–25. <https://doi.org/10.1145/3274301>
- Cheng, Justin; Michael Bernstein; Cristian Danescu-Niculescu-Mizil; and Jure Leskovec (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *CSCW'17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. New York, NY: ACM Press, pp. 1217–1230. <https://doi.org/10.1145/2998181.2998213>
- Clark, Herbert H.; and Susan E. Brennan (1991). Grounding in communication. In L. B. Resnick; J. M. Levine; and S. D. Teasley (eds): *Perspectives on Socially Shared Cognition*. Washington, DC: American Psychological Association, pp. 127–149. <https://doi.org/10.1037/10096-006>
- Clark, Herbert H.; and Edward F. Schaefer (1989). Contributing to discourse. *Cognitive Science*, vol. 13, no. 2, pp. 259–294. https://doi.org/10.1207/s15516709cog1302_7
- Clark, Herbert H.; and Deanna Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition*, vol. 22, no. 1, pp. 1–39. [https://doi.org/10.1016/0010-0277\(86\)90010-7](https://doi.org/10.1016/0010-0277(86)90010-7)
- Coleman, E. Gabriella (2014). *Hacker, Hoaxer, Whistleblower, Spy: The many Faces of Anonymous*. London; New York: Verso.
- Coles, Bryn A.; and Melanie West (2016). Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, vol. 60, pp. 233–244. <https://doi.org/10.1016/j.chb.2016.02.070>
- Condon, Sherri L.; and Claude G. Cech (2001). Profiling turns in interaction: Discourse structure and function. *HICSS'34: Proceedings of the 34th Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Press, pp. 4034–4044. <https://doi.org/10.1109/HICSS.2001.926501>
- Cook, C.; Rianne Conijn; Juliette Schaafsma; and Marjolijn Antheunis (2019). For whom the gamer trolls: A study of trolling interactions in the online gaming context. *Journal of Computer-Mediated Communication*, vol 24, pp. 293–318. <https://doi.org/10.1093/jcmc/zmz014>
- Cruz, Angela Gracia B.; Yuri Seo; and Mathew Rex (2018). Trolling in online communities: A practice-based theoretical perspective. *The Information Society*, vol. 34, no. 1, pp. 15–26. <https://doi.org/10.1080/01972243.2017.1391909>
- Crystal, David 2001. *Language and the Internet*. Cambridge, UK: Cambridge University Press.
- Danet, Brenda (2013). Flaming and linguistic impoliteness on a Listserv. In S. Herring; D. Stein; and T. Virtanen (eds): *Pragmatics of Computer-Mediated Communication*. Berlin, Germany: De Gruyter Mouton, pp. 639–664. <https://doi.org/10.1515/9783110214468.639>

- Dersley, Ian; and Anthony J. Wootton (2000). Complaint sequences within antagonistic argument. *Research on Language and Social Interaction*, vol. 33, no. 4, pp. 375–406. https://doi.org/10.1207/S15327973RLSI3304_02
- Dolata, Mateuz; and Gerhard Schwabe (2019). Translation and adoption: Exploring vocabulary work in expert-layperson encounters. *Computer Supported Cooperative Work (CSCW)*, vol. 28, pp. 685–722. <https://doi.org/10.1007/s10606-019-09358-9>
- Donath, Judith S. (1999). Identity and deception in the virtual community. In M. A. Smith; and P. Kollock (eds): *Communities in Cyberspace*. London and New York: Routledge, pp. 29–59.
- Drew, Paul (1978). Accusations: The occasioned use of members' knowledge of 'religious geography' in describing events. *Sociology*, vol. 12, no. 1, pp. 1–22. <https://doi.org/10.1177/003803857801200102>
- Fichman, Pnina; and Madelyn R. Sanfilippo (2016). *Online Trolling and its Perpetrators: Under the Cyberbridge*. Lanham: Rowman and Littlefield.
- Fleiss, Joseph L. (1981). *Statistical Methods for Rates and Proportions*. Hoboken, NJ: Wiley.
- Garcia, Angela; and Jennifer Baker Jacobs (1998). The interactional organization of computer mediated communication in the college classroom. *Qualitative Sociology*, vol. 21, pp. 299–317.
- Garcia, Angela C.; and Jennifer Baker Jacobs (1999). The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on Language and Social Interaction*, vol. 32, no. 4, pp. 337–367. https://doi.org/10.1207/S15327973rls3204_2
- Garfinkel, Harold (1967). *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.
- Giles, David; Wyke Stommel; Trena M. Paulus; Jessica N. Lester; and Darren Reed (2015). Microanalysis of online data: The methodological development of "digital CA". *Discourse, Context and Media*, vol. 7, pp. 45–51. <https://doi.org/10.1016/j.dcm.2014.12.002>
- Goffman, Erving (1967). *Interaction Ritual: Essays in Face to Face Behavior*. Garden City, NY: Doubleday.
- Golf-Papez, Maja; and Ekant Veer (2017). Don't feed the trolling: Rethinking how online trolling is being defined and combated. *Journal of Marketing Management*, vol. 33, no. 15–16, pp. 1336–1354. <https://doi.org/10.1080/0267257X.2017.1383298>
- Goodwin, Charles; and Marjorie H. Goodwin (1987). Concurrent operations on talk: Notes on the interactive organization of assessments. *IprA Papers in Pragmatics*, vol. 1, no. 1, pp. 1–54. <https://doi.org/10.1075/iprapip.1.1.01goo>
- Graham, Elyse (2019). Boundary maintenance and the origins of trolling. *New Media and Society*, vol. 21, no. 9, pp. 2029–2047. <https://doi.org/10.1177/1461444819837561>
- Greatbatch, David (1986). Some standard uses of supplementary questions in news interviews. In J. Wilson; and B. Crow (eds): *Belfast Working Papers in Language and Linguistics*, vol. 8. Jordanstown, Ireland: University of Ulster, pp. 86–123.
- Grice, H. Paul (1975). Logic and conversation. In P. Cole; and J. L. Morgan (eds): *Syntax and Semantics*, vol. 3, Speech Acts, pp. 41–58. New York, NY: Academic Press.
- Gruber, Helmut (1998). Computer-mediated communication and scholarly discourse: Forms of topic initiation and thematic development. *Pragmatics*, vol. 8, no. 1, pp. 21–47.
- Hankes, Keegan (2015). Black hole. *Southern Poverty Law Center*. <https://www.splcenter.org/fighting-hate/intelligence-report/2015/black-hole>. Accessed 8 November 2019.
- Hård af Segerstad, Ylva (2002). *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. Doctoral Dissertation, Göteborg University, Sweden. <http://nl.ijs.si/janes/wp-content/uploads/2014/09/segerstad02.pdf>. Accessed 7 July 2020.
- Hardaker, Claire (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, Language, Behaviour, Culture*, vol. 6, no. 2, pp. 215–242. <https://doi.org/10.1515/jplr.2010.011>

- Hardaker, Claire (2013). "Uh. . . not to be nitpicky,,,,,but...the past tense of drag is dragged, not drug.": An overview of trolling strategies. *Journal of Language Aggression and Conflict*, vol. 1, no. 1, pp. 58–86. <https://doi.org/10.1075/jlac.1.1.04har>
- Hardaker, Claire (2017). Flaming and trolling. In C. Hoffmann; and W. Bublitz (eds): *Pragmatics of Social Media*. Berlin, Germany: De Gruyter Mouton, pp. 493–522. <https://doi.org/10.1515/9783110431070-018>
- Heritage, John (1984). *Garfinkel and Ethnomethodology*. Cambridge, UK: Polity Press.
- Heritage, John (2012). The epistemic engine: Sequence organization and territories of knowledge. *Research on Language and Social Interaction*, vol. 45, no. 1, pp. 30–52. <https://doi.org/10.1080/08351813.2012.646685>
- Heritage, John; and Geoffrey Raymond (2005). The terms of agreement: Indexing epistemic authority and subordination in talk-in-interaction. *Social Psychology Quarterly*, vol. 68, no. 1, pp. 15–38. <https://doi.org/10.1177/019027250506800103>
- Herring, Susan C. (1994). Politeness in computer culture: Why women thank and men flame. In M. Bucholtz; A.C. Liang; L. A. Sutton; and C. Hines (eds): *Cultural Performances: Proceedings of the Third Berkeley Women and Language Conference*, University of California, pp. 278–294. <http://ella.slis.indiana.edu/~herring/politeness.1994.pdf>
- Herring, Susan C. (1999). Interactional coherence in CMC. *Journal of Computer-Mediated Communication*, vol. 4, no. 4. 10.1111/j.1083-6101.1999.tb00106.x
- Herring, Susan C. (2010). Computer-mediated conversation: Introduction and overview. *Language@Internet*, vol. 7, article 2. <https://www.languageatinternet.org/articles/2010/2801/?searchterm=Herring>. Accessed 7 July 2020.
- Herring, Susan; Kirk Job-Sluder; Rebecca Scheckler; and Sasha Barab (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society*, vol. 18, no. 5, pp. 371–384. <https://doi.org/10.1080/01972240290108186>
- Hopkinson, Christopher (2013). Trolling in online discussions: From provocation to community-building. *Brno Studies in English*, vol. 39, no. 1, pp. 5–25.
- Hopper, Robert (2005). A cognitive agnostic in conversation analysis: When do strategies affect spoken interaction? In H. te Molder; and J. Potter (eds.): *Conversation and Cognition*. Cambridge, UK: Cambridge University Press, pp. 134–158. <https://doi.org/10.1017/CBO9780511489990.007>
- Humphrecht, Edda (2018). Where 'fake news' flourishes: A comparison across four western democracies. *Information, Communication and Society*, vol. 22, no. 13, pp. 1973–1988. <https://doi.org/10.1080/1369118X.2018.1474241>
- Kirman, Ben; Conor Lineham; and Shaun Lawson. (2012). Exploring mischief and mayhem in social computing or: How we learned to stop worrying and love the trolls. *CHI EA '12: CHI '12 Extended Abstracts on Human Factors in Computing Systems*. New York, NY: ACM Press, pp. 121–130. <https://doi.org/10.1145/2212776.2212790>
- Knustad, Magnus (2020). Get lost, troll: How accusations of trolling in newspaper comment sections affect the debate. *First Monday*, vol. 25, no. 8. <https://doi.org/10.5210/fm.v25i8.10270>
- Krauss, Robert M.; and Susan R. Fussell (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, vol. 9, no. 1, pp. 2–24. <https://doi.org/10.1521/soco.1991.9.1.2>
- Levinson, Stephen C. (1983). *Pragmatics*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511813313>
- Lewis, Helen (2015). Sexist, Racist – The Web Hounding of Ellen Pao Shows the Trolls Are Winning. *The Guardian*. <https://www.theguardian.com/commentisfree/2015/jul/17/ellen-pao-reddit-sexist-racist-internet-trolls-winning>. Accessed 8 November 2019.
- Lindholm, Loukia (2013). The maxims of online nicknames. In S. Herring; D. Stein; and Tuija Virtanen (eds): *Pragmatics of Computer-Mediated Communication*. Berlin, Germany: De Gruyter Mouton, pp. 437–462. <https://doi.org/10.1515/9783110214468.437>

- Lumsden, Karen; and Heather Morgan (2017). Media framing of trolling and online abuse: silencing strategies, symbolic violence, and victim blaming. *Feminist Media Studies*, vol. 17, no. 6, pp. 926–940. <https://doi.org/10.1080/14680777.2017.1316755>
- Markman, Kris M. (2005). To send or not to send: Turn construction in computer-mediated chat. *Texas Linguistic Forum*, vol. 48, pp. 115–124.
- Markman, Kris M. (2010). Learning to work virtually: Conversational repair as a resource for norm development in computer-mediated team meetings. In J. Park; and E. Abels (eds): *Interpersonal Relations and Social Patterns in Communication Technologies: Discourse Norms, Language Structures and Cultural Variables*. Hershey, PA: IGI Global, pp. 220–236.
- Meredith, Joanne (2019). Conversation analysis and online interaction. *Research on Language and Social Interaction*, vol. 52, no. 3, pp. 241–256. <https://doi.org/10.1080/08351813.2019.1631040>
- Meredith, Joanne; and Elizabeth Stokoe (2014). Repair: Comparing Facebook ‘chat’ with spoken interaction. *Discourse and Communication*, vol. 8, no. 2, pp. 181–207. <https://doi.org/10.1177/1750481313510815>
- Moore, R. J.; and Raphael Arar (2019). *Conversational UX Design: A Practitioner’s Guide to the Natural Conversation Framework*. New York, NY: ACM Press. <https://doi.org/10.1145/3304087>
- Moore, Robert J.; and Elizabeth F. Churchill (2011). Computer interaction analysis: Toward an empirical approach to understanding user practice and eye gaze in GUI-based interaction. *Computer Supported Cooperative Work (CSCW)*, vol. 20, pp. 497–528. <https://doi.org/10.1007/s10606-011-9142-2>
- Moore, Robert J.; Nicolas Ducheneaut; and Eric Nickell (2007). Doing virtually nothing: Awareness and accountability in massively multiplayer online worlds. *Computer Supported Cooperative Work (CSCW)*, vol. 16, pp. 265–305. <https://doi.org/10.1007/s10606-006-9021-4>
- Pew Research Center (2016). Nearly Eight-in-Ten Reddit Users Get News on the Site. https://www.pewresearch.org/wp-content/uploads/sites/8/2016/02/PJ_2016.02.25_Reddit_FINAL.pdf. Accessed 8 November 2019.
- Phillips, Whitney (2011). LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online. *First Monday*, vol. 16, no. 12.
- Phillips, Whitney (2015). *This is Why We Can’t Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture*. Cambridge, MA: The MIT Press.
- Pomerantz, Anita (1984). Pursuing a response. In J. M. Atkinson; and J. Heritage (eds): *Structures of Social Action*. Cambridge, UK: Cambridge University Press, pp. 152–164.
- Rintel, E. Sean; and Jeffery Pittam, J. (1997). Strangers in a strange land: Interaction management on Internet relay chat. *Human Communication Research*, vol. 23, no. 4, pp. 507–534. <https://doi.org/10.1111/j.1468-2958.1997.tb00408.x>
- Romaniuk, Tanya (2013). Pursuing answers to questions in broadcast journalism. *Research on Language and Social Interaction*, vol. 46, no. 2, pp. 144–164. <https://doi.org/10.1080/08351813.2013.780339>
- Ruhleder, Karen; and Brigitte Jordan (2001). Co-constructing non-mutual realities: Delay-generated trouble in distributed interaction. *Computer Supported Cooperative Work (CSCW)*, vol. 10, pp. 113–138. <https://doi.org/10.1023/A:1011243905593>
- Sacks, Harvey (1992). *Lectures on Conversation*, vol. 1 (Fall 1964–Spring 1968). Oxford: Blackwell.
- Sanfilippo, Madelyn; Shengnan Yang; and Pnina Fichman (2017a). Trolling Here, There, and Everywhere: Perceptions of Trolling Behaviors in Context. *Journal of the Association for Information Science and Technology*, vol. 68, no. 10, pp. 2313–2327. <https://doi.org/10.1002/asi.23902>
- Sanfilippo, Madelyn; Shengnan Yang; and Pnina Fichman (2017b). Managing online trolling: From deviant to social and political trolls. In *HICSS’50: Proceedings of the 50th Hawaii International Conference on System Sciences*. University of Hawai’i at Manoa. <https://doi.org/10.24251/HICSS.2017.219>

- Sanfilippo, Madelyn R.; Pnina Fichman; and Shengnan Yang (2018). Multidimensionality of online trolling behaviors. *The Information Society*, vol. 34, no. 1, pp. 27–39. <https://doi.org/10.1080/01972243.2017.1391911>
- Schegloff, Emanuel A. (1968). Sequencing in conversational openings. *American Anthropologist*, vol. 70, no. 6, pp. 1075–1095. <https://doi.org/10.1525/aa.1968.70.6.02a00030>
- Schegloff, Emanuel A. (2007). *Sequence Organization in Interaction: A Primer in Conversation Analysis*. Cambridge, UK: Cambridge University Press.
- Schegloff, Emanuel A.; and Harvey Sacks (1973). Opening up closings. *Semiotica*, vol. 8, no. 4. <https://doi.org/10.1515/semi.1973.8.4.289>
- Schegloff, Emanuel A.; Gail Jefferson; and Harvey Sacks (1977). The preference for self-correction in the organization of repair in conversation. *Language*, vol. 53, no. 2, pp. 361–382.
- Shachaf, Pnina; and Noriko Hara (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, vol. 36, no. 3, pp. 357–370. <https://doi.org/10.1177/0165551510365390>
- Skovholt, Karianne; and Jan Svennevig (2013). 24. Responses and non-responses in workplace emails. In S. Herring; D. Stein; and T. Virtanen (eds): *Pragmatics of computer-mediated communication*, vol. 9. Walter de Gruyter. pp. 589–612.
- Smith, Mark; J. J. Cadiz; and Byron Burkhalter (2000). Conversation trees and threaded chats. In *CSCW'00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*. New York, NY: ACM Press, pp. 97–105. <https://doi.org/10.1145/358916.358980>
- Stivers, Tanya (2013). Sequence organization. In J. Sidnell; and T. Stivers (eds): *The Handbook of Conversation Analysis*, First Edition. Blackwell Publishing Ltd. pp. 191–209.
- Stivers, Tanya; and Makoto Hayashi (2010). Transformative answers: One way to resist a question's constraints. *Language in Society*, vol. 39, no. 1, pp. 1–25. <https://doi.org/10.1017/S0047404509990637>
- Stivers, Tanya; and Jeffrey D. Robinson (2006). A Preference for progressivity in interaction. *Language in Society*, vol. 35, no. 3, pp. 367–392. <https://doi.org/10.1017/S0047404506060179>
- Stivers, Tanya; and Federico Rossano (2010). Mobilizing response. *Research on Language and Social Interaction*, vol. 43, no. 1, pp. 3–31. <https://doi.org/10.1080/08351810903471258>
- Stivers, Tanya; Nick J. Enfield; and Stephen C. Levinson (2010). Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, vol. 42, no. 10, pp. 2615–2619. <https://doi.org/10.1016/j.pragma.2010.04.001>
- Stommel, Wyke; and Tom Koole (2010). The online support group as a community: A micro-analysis of the interaction with a new member. *Discourse Studies*, vol. 12, no. 3, pp. 357–378. <https://doi.org/10.1177/1461445609358518>
- Sudweeks, Fay; and Shezaf Rafaei (1996). How do you get a hundred strangers to agree? Computer mediated communication and collaboration. In T. Harrison; and T. Stephens (eds): *Computer Networking and Scholarly Communication in the Twenty-First-Century University*, pp. 115–136. Albany, NY: SUNY Press.
- Tepper, Michele (1997). Usenet communities and the cultural politics of information. In D. Porter (ed.): *Internet Culture*. New York, NY: Routledge, pp. 39–54.
- Thompson, Sandra A.; Barbara A. Fox; and Elizabeth Couper-Kuhlen (2015). *Grammar in Everyday Talk: Building Responsive Actions*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9781139381154>
- Turner, Tammara C.; Marc A. Smith; Danyel Fisher; and Howard T. Welser (2005). Picturing Usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, vol. 10, no. 4. <https://doi.org/10.1111/j.1083-6101.2005.tb00270.x>
- Twitchell, Douglas P.; and Jay F. Nunamaker (2004). Speech act profiling: A probabilistic method for analyzing persistent conversations and their participants. In *HICSS'37: Proceedings of the 37th Annual Hawaii International Conference on System Sciences*. New York, NY: IEEE, pp. 1713–1722. <https://doi.org/10.1109/HICSS.2004.1265283>

- Vatanen, Anna (2014). *Responding in Overlap: Agency, Epistemicity and Social Action in Conversation*. Ph.D. dissertation. Helsinki, Finland: University of Helsinki.
- Wagner, Christian; Rachael K.F. Ip; Karen S. K. Cheung; and Fion S. L. Lee (2005). Deceptive communication in virtual communities. In *HICSS'38: Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. New York, NY: IEEE, pp. 1–8. <https://doi.org/10.1109/HICSS.2005.185>
- Williams, Hywel T. P.; James R. McMurray; Tim Kurz; and F. Hugo Lambert (2015). Network Analysis Reveals Open Forums and Echo Chambers in Social Media Discussions of Climate Change. *Global Environmental Change*, vol. 32, pp. 126–138. <https://doi.org/10.1016/j.gloenvcha.2015.03.006>
- Woodruff, Allison and Paul M. Aoki (2004). Push-to-talk social talk. *Computer Supported Cooperative Work (CSCW)*, vol. 13, pp. 409–441. <https://doi.org/10.1007/s10606-004-5060-x>
- Zelenkauskaitė, Asta; and Niezgodą, Brandon (2017). “Stop Kremlin trolls:” Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting. *First Monday*, vol. 22, no. 5. <https://doi.org/10.5210/fm.v22i5.7795>
- Zvereva, Vera (2020). Trolling as a digital literary practice in the Russian language Internet. *Russian Literature*, vol. 118, pp. 107–140. <https://doi.org/10.1016/j.ruslit.2020.11.005>.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.