LIMITS OF HATE SPEECH AND FREEDOM OF SPEECH ON MODERATED NEWS WEBSITES IN FINLAND, SWEDEN, THE NETHERLANDS AND THE UK

Reeta PÖYHTÄRI
University of Tampere, School of Communication, Media and Theatre, Research Centre for Journalism, Media and Communication (Comet), Kalevantie 4/E313, 33014, Finland
e-mail: reeta.poyhtari@uta.fi

ABSTRACT

The article discusses and compares the structure and moderation practices of comment fields on news websites in Finland, Sweden, the Netherlands and Great Britain. The focus is on the tension between freedom of speech and moderation of hate speech and cyberhate. The research data is comprised of interviews with 16 moderators along with analysis of comment field structure and guidelines for moderation practices from 18 news websites. The news media actively prevents hate speech and cyberhate in ways that differ by country. Nonetheless, hate speech and cyberhate on news websites is similar in all countries and moderated based on three types of regulations: laws, media ethics and self-regulatory guidelines.

Keywords: news comments, freedom of speech, hate speech, cyberhate, moderation.

I LIMITI DEL DISCORSO INCITANTE ALL'ODIO E LA LIBERTA' DI PAROLA IN SITI DI NEWS CON MODERAZIONE IN FINLANDIA, SVEZIA, PAESI BASSI, E GRAN BRETAGNA

SINTESI

Nell' articolo sono discusse e comparate le strutture e le prassi di moderazione negli spazi per i commenti nei siti web di notizie in Finlandia, Svezia, Paesi Bassi e Gran Bretagna. Particolare attenzione è rivolta alla tensione tra la libertà di parola e la moderazione degli hate speech (messaggi di istigazione all'odio) e del cyberhate (cyberodio). I dati della ricerca consistono in interviste effettuate a 16 moderatori e nell'analisi delle strutture del campo per commenti e indicazioni per le prassi di moderazione presenti in 18 siti web di notizie. La stampa si impegna attivamente a combattere gli hate speech e il cyberhate sui siti web di notizie in modi che variano tra i diversi Paesi. Tuttavia gli hate speech e il cyberhate nei siti web di notizie sono simili in tutti i Paesi e vengono moderati con tre tipi di regolamento: leggi, etica del web e codici di autoregolamentazione.

Parole chiave: Commenti alle notizie, libertà di parola, hate speech, cyberhate, moderazione.

INTRODUCTION

Along with discussion forums in social media, comment sections on online news sites have become popular. In order to finance the news business, news media have thought of new ways to attract and engage readers and make their visits to news sites longer. Through that, the news media hope to gain advertisers. News comment sections also suit the news media's ideal of being enablers of public debate and democracy. Comment fields would potentially add value to a news site by enhancing public's participation, allowing contact between the media and its users, and helping the news outlet to express public's views (e.g., Deuze, 2003; Deuze et al., 2007; Domingo et al., 2008; Erjavec & Poler Kovačič, 2012).

After the initial enthusiasm, which in many cases left the public free to comment all news material on a news site, the news media have had to adjust their services to the fact that a substantial part of user generated content does not meet the standards to be published. The comment function generates a great deal of extra work for editors, as the comments need to be checked and moderated. News comment fields are further problematic, as they besides more innocent material include hate speech and other aggressive, insulting and stigmatizing content, known as cyberhate (Hughey & Daniels, 2013). This creates another dilemma for the news media. For media sites maintaining news comment fields, the great variety of potential hate speech and cyberhate content necessitates awareness and sensitivity to a plethora of verbal misbehaviours that can be and are offensive to various groups, minorities and individuals in society. At the same time, the media seeks to remain alert to its core value as a defender of freedom of speech and enabler of public debate, also on sensitive issues such as inter-ethnic relations. Media has to balance between providing public an access to free speech, while guaranteeing that such debate is conducted responsibly and ethically.

Earlier research on the issue of comment fields on news sites has discussed the journalistic value and use of news comments (Deuze, 2006; Heinonen, 2011), the commonalities between the practice of hosting public discourse on news sites and on social media platforms (Braun & Gillespie, 2011), the practices and difficulties of moderation of news comments (Trygg, 2012), as well as the contents of news comment fields (e.g., Canter, 2013). In relation to the question of cyberhate, especially the contents of racist news and blog comments have been covered (Cammaerts, 2009; Horsti & Nikunen, 2013) and it has been discussed, whether moderated news comments affect the forms of verbal racism used in discussions (Hughey & Daniels, 2013).

Previous research has mainly been national, comparing various news sites within one country (Hermida & Thurman, 2008), while international comparisons have been rare (Trygg, 2012; Goodman, 2013). Research has often approached user comments from the point of view of journalists, while in fact nowadays typically specialized moderators face the comments and make decisions concerning publishing (see e.g., Trygg, 2012). It has also not been widely studied, how news media organisations exactly recognize hate speech and cyberhate when moderating and how they make the decisions to remove or publish user comments (for an on-going study, see Benesch, 2013). It has also not been discussed in specific, how the news media perceive the controversy between allowing free speech on sensitive societal issues, while also protecting the public from racism, hate speech and cyberhate.

This article attempts to fill in this gap in research and discusses the structure of news media's comment sections as well as their moderation practices and regulations concerning cyberhate and hate speech. The article asks, with what kind of solutions the news media on the one hand enable free public debate in news comment fields in accordance with their aims, and on the other hand restrict it to avoid problematic contents, such as hate speech, in advance. This analysis is done by studying the structure of the comment fields of 18 news comment fields. Secondly, it is asked, what do the moderators of various news comment sites recognise as hate speech, cyberhate or other vice problematic content? How do moderators make the decisions to publish or to remove such contents? In order to analyse this, the moderators of 16 news comment fields were interviewed and the user and moderation guidelines of those comment fields were analysed. Thirdly, it is asked what the possible outcomes of these practices in relation to prevention of hate speech, but also to promotion of free speech and user participation could be, and which values of the news media are reflected. The article considers news sites in Finland, Sweden, the Netherlands and the UK.

THEORETICAL FRAMEWORK

The news media have adjusted to the development of Web 2.0 and created their own services for public debate. At the same time they have come to face the ill-sides of Internet discussions, including hate speech and cyberhate.

Freedom of speech and public debate on news comment sites

Thanks to the Internet, and especially the development of Web 2.0, possibilities for public debate and deliberation have increased exponentially. The new web enables users to spread all kinds of user generated content to unlimited publics. On the surface, freedom of speech is wider than ever (e.g., Sunstein, 2001; Margolis & Moreno-Riaño, 2009).

However, realisation of true freedom of speech on the Internet encounters serious problems, digital divide

to start with (Jørgensen, 2013). In addition, free public deliberation between various people and opinions remains a dream, as typically issue- and audience specific Internet platforms regularly turn into so-called echo-chambers. The Internet and social media on the whole leave people free to determine the content they want to consume, as well as the company they'd like to keep (e.g., Sunstein, 2001; Papacharissi, 2002; Youngs, 2009). People's behaviour on the Internet is strongly connected to their affects, so they seek appropriate platforms with like-minded individuals, serving their personal interests (Tsesis, 2002; Douglas, 2008; Joinson et al., 2008). As a result, people are not likely to encounter opinions different to their own.

Unlike to subject specific discussion forums, comment fields of the news media have not been created to please a very specific group of audience only. Like print newspapers, they ought to serve the public interest at large. Freedom of speech, public participation and democracy are valued as ideals when the news media design and set up their online news comment sections (Deuze, 2003; Deuze et al., 2007; Domingo et al., 2008). By enabling user comments, users are given a chance for free debate and to encounter various views. At the same time, the media gains useful insights into public opinion. In local settings, especially, such discussions can create a bond between the media and the public (Heinonen 2011; Canter 2013). For these reasons, news comment fields are a fora where a variety of people and opinions have a chance to meet, and where actual deliberation potentially could be practiced.

The ideal of the news media as enablers of public deliberation and democracy has its roots in the classical liberal understandings of free speech (esp. John Stuart Mill), as well as in the models of deliberative democracy (e.g., Habermas, 1984). Ideally, free discussion allows for finding the truth (Mill, 1982). It is the news media's role to support the public's opportunities for self-expression and truth-seeking by offering a public sphere for deliberation (e.g., Nieminen & Nordenstreng, 2012).

The liberal ideals of freedom of speech and deliberation however include a normative element. It often has been claimed mistakenly that John Stuart Mill was the inventor or true supporter of the 'free marketplace of ideas', allowing public expressions of all opinions (Nordenstreng, 2013). Instead, freedom of speech, as discussed by Mill and Habermas, applies ideally only to those who are well-informed and civilized enough to deliberate responsibly on public and common issues. It is not to be used by everyone for all possible purposes. Traditionally the news media has controlled the participants of public deliberation by deciding, whose opinion is to be heard in the public debate in the news material, but also in letters-to-editors.

Many have criticised the theoretical ideas of both Mill and Habermas that limit the group of possible participants in public debates (e.g., Young, 2000; Downey

& Fenton, 2003; Mouffe, 2005; Fraser, 2007). These ideas simply do not seem applicable in today's Web 2.0 environments where the official and normatively limited public spheres provided by e.g. news media are competing with endless alternative, and normatively less restricted platforms for people's deliberation activities. Not surprisingly, the news media have encountered difficulties, as their comment sections have been overwhelmed with uncivilized and unruly contents. Not all participants are willing to keep themselves to the ideal of expressing only civilized and well-informed comments. News media have found themselves hosting an unruly public (Braun & Gillespie, 2011; da Silva, 2013).

Regrettably, the comment sections of news media have also not remained free of hate speech and other common forms of Internet cyberhate (Hughey & Daniels, 2009). People active on actual hate sites and other fora of social media are keen to make visits to other discussions where their message can be spread to the wider public. In particular, anonymous news comment fields tempt such individuals to leave hateful contributions to debates (Back, 2002; Cooper, 2004; Roversi, 2008; Cammaerts, 2008 & 2009; Daniels, 2009).

Moderation of news comments, hate speech and cyberhate

In order to host the public debate in news comment fields, news media have of late set up a variety of enhanced moderation practices to control discussions. The news media have a legal and ethical obligation to function responsibly and news comment fields fall under that responsibility. Moderation is a grassroots-level solution to control website content. Comments are removed either in advance (pre-moderated), which means that they are not published at all, or afterwards (post-moderated), if they appear potentially illegal or inappropriate after publishing. Content sometimes can be edited or re-written before publishing. It should be noted that moderation is not the same as censorship, which only can be performed by states (Hannula & Neuvonen, 2011).

In the moderation practice, the great variety of potential hate speech and cyberhate content necessitates awareness and sensitivity to a plethora of verbal misbehaviours. At the same time, the media seeks to remain alert to its core value as a defender of free speech and enabler of public debate and democracy. This poses the news media a dilemma: it has to balance between providing public access to free speech, while guaranteeing that such debate is conducted responsibly.

The dilemma that the news media is faced with, also flows from legislation that is not unambiguous in definition of hate speech. Freedom of speech functions as a starting point, and it is guaranteed in the legislations of modern democracies and by international conventions. In Europe, the European Convention of Human Rights (1950) and the related Court of Human Rights (ECtHR)

protect the exercise of these rights. The news media use the same rights, when they publish contents created by journalists or the public. In relation to hate speech ECtHR has in its jurisprudence expressed that besides positive, harmless and insignificant expressions freedom of speech also covers expressions that can be worrying, shocking or insulting, or include provocation (Weber, 2009). Which utterances are covered by freedom of speech, is however not easy to judge.

Generally, there are two alternative ways to consider free speech and its limitations in relation to hate speech. Firstly, freedom of speech is seen as an ultimate freedom that individuals should be able to exercise without restrictions. Restrictions are only justifiable in the rare case where direct harm is caused to others. Some have questioned whether speech is an actual act at all, and whether it has the capacity to cause any true or direct harms. What can be understood as 'direct harm' is a complicated question, and therefore, this 'harm principle' is difficult to enact in practice. However, unlimited free speech and the harm principle have gained support from free speech proponents who argue for as few restrictions as possible. Legislation in the United States follows this line of thought and grants a high value on free speech with minimum restrictions (e.g., Calvert, 1997; Tsesis, 2002; Bleich, 2011; Rønning, 2013).

The other tradition sees speech as an act that can cause various sorts of damages, direct and indirect, not only to its recipients but also to society on the whole. Speech is taken as a powerful tool that can cause long-term harm to minorities by marking them as subordinate or inferior, which can lead to general public unrest and hostility between groups (Calvert, 1997; Tsesis, 2002). Free speech is defined as a fundamental right, which must be protected by law from censorship, but which may not be used to endanger other fundamental rights of other individuals or groups, including their human dignity. Various rights need to be balanced against each other (Kortteinen, 1996). This view also has encountered critique: If certain groups, typically minorities, are to be protected from verbal offences, who is in the position to define which groups count as such minorities and what counts as a penal offence (Molnar, 2012; Rønning, 2013)? Despite the difficulties in practice, most European legislations today recognize the need to protect vulnerable groups from harmful and offensive uses of free speech.

The term 'hate speech' has rarely been defined as such in national legislations. Rather, many nations have criminalized e.g. racist speech and incitement to racial, ethnic or religious hatred; discrimination, provocation and defamation (Bleich, 2011). The definitions of these crimes vary by country, and as a result, hate speech and hate crimes have slightly different meanings (Garland & Chakraborti, 2012). The most specific definition to hate speech has been given by The Council of Europe's Committee of Ministers' (1997), articulating it as follows: 'the term "hate speech" shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.'

The Internet is a platform not only for hate speech, but also for other hateful expressions. An overarching term that has been used for this kind of content on the web is cyberhate, which has been defined as 'the use of any electronic technology to spread bigoted, discriminatory, terrorist and extremist information' (Edelstein & Wolf, 2013, 3). More specifically, cyberhate includes a broad selection of behaviours and expressions: racism, anti-Semitism, religious bigotry, homophobia, bigotry aimed at the disabled, political hatred, rumour-mongering, misogyny and violent pornography, promotion of terrorism, cyberbullying, harassment and stalking, speech that silences counter-speech such as slurs, insults and epithets and speech that defames an entire group (ibid., 8). Clearly, cyberhate as a term is broader than 'hate speech' defined in criminal laws. The scope of utterances that cause trouble to news comment fields, is thus much wider than hate speech only, and it poses a challenge to the media that moderates discussions.

Besides the actual contents of news comments, profitability is another vital element affecting the moderation practices of the news media. The comment section should be an economically beneficial part of the media product (Hughey & Daniels, 2013). In many cases, profitability has proven difficult, as the maintenance and development of the platforms is expensive and laborious while exact benefits remain vague. Comments should add to the quality, not reduce it. It is a true challenge for the media to produce qualitatively desirable discussions using as few resources as possible. Ultimately, profitability affects discussion structure, moderation resources and implemented practices (see also Trygg, 2012).

Taken together, there are several factors creating a quagmire within which moderation occurs: the media's ambition and ideal to create spaces for public debate, the need to protect free speech but to do it qualitatively well, the responsibility to eliminate hate speech and other misbehaviours, and the necessity of profitability of business. With this complexity in mind, it is relevant to inquire how media have approached these dilemmas.

The key research questions to this article can be put as follows: How do various news media in Finland, Sweden, the Netherlands and the UK treat the dilemma of allowing public debate, but to avoid hate speech? What kind of news comment field structures have various news media developed in order to on the one hand enhance public discussion, and to avoid inappropriate comments on the other? When moderating, how do moderators decide what constitutes hate speech or cyberhate in need of moderation, as the laws are ambiguous?

Taking into account the complicated field within which the media performs their moderation practices, it is relevant to ask about the limits between hate speech and free speech. What values do the chosen moderation practices in various media reflect? The question also arises whether moderated discussions enable public participation, or whether it limits participants' freedom of speech. What are the possible consequences of moderation of news comments?

METHODS

For this research, we first of all analysed the structure of in total 18 comment fields of news sites in Finland (8), the Netherlands (4), Sweden (3) and the UK (3). The studied news media were Helsingin Sanomat, Iltalehti, Ilkka, Kainuun Sanomat, Uusi Suomi, Yle, Aamulehti and Satakunnan kansa in Finland; De Volkskrant, Algemeen Dagblad, Trouw and De Telegraaf in the Netherlands; Dagens Nyheter, Aftonbladet and Sveriges radio in Sweden; and Guardian, Daily Mail and The Telegraph in Great Britain. For the Finnish material the aim was to include various media houses and types of news outlets in the study. We selected a national, local (4), a tabloid, a public broadcaster and a web-based only news media. We also paid attention to the popularity of the comment fields when making the selection. In the selection for the Netherlands, Sweden and Great Britain we included a quality and a tabloid-like news outlet with the most popular comment fields, and since the Swedish public broadcasting service closely resembles the Finnish one, it was included as well. The selection for the Netherlands covers three different quality newspapers of one publisher, as their moderation is taken care of jointly.

The qualitative analysis of comment field structure was performed to find out, how the comment fields enable public debate, but also restrict problematic contents. Besides written user guidelines, the actual technology or structure of the discussion platform steers the discussions (see Braun & Gillespie, 2011; Goodman, 2013). Therefore, it is relevant to analyse the structure of the comment fields. We paid attention to whether and how a user needed to register for the site in order to comment, what news one could comment on, whether moderation was pre- or post-moderation, whether moderators or journalists participated in discussions, and the location of user guidelines on the site. We also considered the core content of the user guidelines: what was allowed/forbidden in comments, whether the public could report problematic contents, whether references were made to any laws, and who was announced to hold responsibility for the comments.

In addition to analysis of the comment field structure and user guidelines, 16 moderators from the studied news sites were interviewed in 2012. In-depth interviews are a much used method in media research. They are an adequate method for the study of organisations and practices (Jankowski & Wester, 1991), and in this case particularly the news organisations' moderation practices. We interviewed moderators or heads of moderation teams of all the selected news media that were willing to give us an interview, one person per each media[1]. In Great Britain, the news media did not want to provide interviews, except for the Guardian. In Sweden, we also tried to interview the moderation company moderating the most news sites, but they refused interview. In the Netherlands instead, we were besides moderators at news media, allowed to interview the moderation company that in practice moderates the most news sites. The interviews are specified in the list of sources.

During the semi-structured in-depth interviews, questions were asked about the structure of the news comment sections and the interactive services for the public; the moderation practices of the news comments and the rules of moderation; and views of the moderators on hate speech, freedom of speech and its restrictions. Also, the internal moderation guidelines were studied, if they were provided to us upon request. In total, some 23 hours of interview material in Finnish, Swedish and English was recorded and transcribed. The transcriptions were analysed qualitatively by categorising the material into various classes and subclasses according to spoken themes and issues, which then were analysed further theoretically. A qualitative method was followed in order to explore moderation practices in different countries and news media, and to understand the variety of practices, but not to produce quantitative data about them.

RESULTS

News comment fields have been set up in the hope that they would create a free debate with an added value to a news site. At start, much or all of the news material was open for commenting, but that lead into too much moderation work. Today, the news media use two strategies in order to avoid problematic contents to be published in their comment fields. First, the structure and the technology of the comment field itself is designed to prevent problems. Secondly, the comments are post- or premoderated according the rules of moderation.

Structure of news comment fields: Allowing debate, preventing hate speech and cyberhate

News media have developed the structure and technology of the comment fields to minimise the need for moderation, and enabling adequate control of the dis-

---

1 Except for Helsingin Sanomat and Kainuun Sanomat, where they preferred to have two interviewees due to recent changes in their moderation team. In Britain, the Guardian gave two separate interviews.

Table 1: News comment field structures in different countries:

News comment field structures	What can people comment on?	Registration required/ not	Moderation (pre/post, performed by)	Moderators participate in discussions/not
Finland	All news on the site	Yes, with user name (or full name); sometimes no registration	Pre-moderation mostly, moderators in the editorial office	Mostly not, sometimes moderator-journalists or journalists participate
The UK	All news on the site	Yes, with user name (or full name, e.g. Facebook (FB) account); sometimes no registration	Post-moderation, moderators in the editorial office	Sometimes; journalists can also participate
The Netherlands	Selected discussion articles and news	Yes, with user name (or full name, e.g. FB account)	Pre-moderation, moderators from an external company	No; journalists can participate sometimes
Sweden	Selected discussion articles, live chats	Yes, with user name (or full name, e.g. FB account)	Post-moderation, moderators from an external company	No; Live chats and special discussions are organised by journalists

cussions with the resources available (see also Goodman, 2013). The chosen practices share commonalities within the countries, but since they are connected to the legal and media ethical framework in each country, variations exist. The table above summarises the general structure of the comment fields in each studied country.

The commenting practices on Finnish news websites are quite liberal. The studied news sites allow all or almost all news to be commented on. To comment, users need to register with a user name, but in tabloid Iltalehti, there are still news discussions in which registration is not needed. The comments are pre-moderated for all the studied news sites, except for Iltalehti and online news media Uusi Suomi, which post-moderate due to volume- and resource-related reasons (interviews, 10/2012).

The ethical code for Finnish journalists (Julkisen sanan neuvosto, 2014) recommends that news media moderate in advance, but the treatment of the law and the ethical code is the same, regardless of whether one pre- or post-moderates. In both cases, the news media can be held responsible for illegal content in news comments, once they have become aware of such content. According to the ethical code, media also has an active responsibility to avoid such content from being published. For all interviewed cases, the moderators work in-house, and they are trained within the news media.

The Finnish practices have much in common with UK practices, where all or most news on the news sites can be commented on. There is variation in terms of whether people need to register or not; most sites require registration with user names. The British comment sections are mainly post-moderated for legal responsi-

bility reasons; often, according to the interviews (Guardian, 1/2013), moderation is performed retroactively as a reaction to user reports of illegal content. An in-house moderating team is used at the Guardian. The team has active contact with both the editorial office as well as the legal department when making decisions concerning moderation. The Guardian trusts in active user engagement and peer-to-peer control, as well as active journalist and expert participation in the discussions when possible. For these reasons, the size of the moderator community team can be kept small despite the large commenting volume.

The practices in the Netherlands differ from those in Finland and in the UK. Instead of opening all news for comments, the Dutch news sites select a small portion of news items that can be commented on daily. The selection, its size and diversity varies per medium, even within one media house (e.g. within the media house De Persgroep, De Volksrant, Algemeen Dagblad and Trouw all follow different policies). It is clear that some news sites allow more discussion than others. Users can comment on news once they have registered.

The moderation of the news sites in the Netherlands is performed by a specified moderation company selling its services to news organisations. Interestingly, the moderation company offers its services to several of the biggest media companies and news sites, which means that news comments are moderated in a standardized way. The moderators are trained by the company itself, and according to their interview (11/2012), the training concentrates on recognition of hate speech and other possibly problematic content as well as speed and ac-

curacy in moderation. Dutch news sites prefer to pre-moderate.

In Sweden, news comments were initially very open for public discussion, but the 2010 parliamentary elections brought a new extreme right-wing political party into the public debate. This caused a significant increase in the number of racist and sexist news comments, attacking both the Swedish immigrant population as well as the basic ideas of racial and gender equality held dear in Swedish society (interviews with Dagens Nyheter 10/2012 and Aftonbladet 1/2013). The news media debated the problematic news comments extensively, and, as a result, many of them decided to allow less discussion on their websites.

Today, the interviewed print media (Dagens Nyheter, Aftonbladet) only open a small selection of news for commenting on a daily basis. If the media wants 'serious' discussions with the publics, live chats hosted by journalists are organised, and preferably, journalists participate in them themselves. The news media prefers to use various social media applications to allow people to share the news; actual comment fields at the end of a news piece have become a rarity. In 2012, Sveriges Radio still let each channel and programme decide on their commenting practices (interview 05/2012); now in 2014, it seems to have followed other news media, mainly allowing sharing through Facebook, Twitter and other services but no comments.

At the Swedish news sites, moderation is taken care of retroactively due to Swedish legislation making the publisher's responsibility greater for advance moderation. When post-moderating, the media are only responsible for removing illegal contents once they are noticed. Similar to the situation in the Netherlands, moderation of the largest news sites is performed by a specialised company that sells their service to news media. These moderators mostly remove illegal content. As was told in an interview (Dagens Nyheter 10/2012), moderators receive standard training, focused especially on media law and ethics, organised by the Swedish Media Institution (Fojo).

It is with and within these structures that various news sites in the studied countries are trying to prevent hate speech and cyberhate from being published. In this way, the media is making decisions concerning people's possibilities to express themselves freely in the news comments. All the solutions chosen enable prevention of hate speech and cyberhate, but some solutions allow the exercise of free speech more extensively than others. The Finnish and British cases seem quite liberal, the post-moderated British media being the most permissive for free debate. The Swedish and Dutch cases, with their selected news discussions, are more restrictive. The Dutch news sites allow more discussion than the Swedish ones, but because they are pre-moderated,

the Dutch discussions are possibly the most restricted. The structure of the news comment fields, however, is only the first layer of practices implemented by the media in the overall struggle against cyberhate. In the end, hate speech and other problematic contents are removed from the discussions according to the actual moderation regulations and guidelines, which vary by media and country.

Moderation of hate speech and cyberhate: Three types of moderation regulation

Hate speech, including ethnic and racial hatred, was familiar to the interviewed moderators in all countries. In addition, moderators encountered hatred towards women, political hatred and hatred directed to individuals, either a public person or a regular user of a comment field. Moderators thus face all forms of cyberhate discussed above (Edelstein & Wolf, 2013). The regulations that direct moderation also share similarities in all countries studied.

In the moderation practice, there are three types of regulations used by media when making decisions concerning the news comments. First of all, like the guidelines given to users by the news sites often already indicate, the comments need to comply with local laws. In some guidelines, like those of the Dutch De Telegraaf, direct links are provided to the laws concerning incitement to hatred towards various minorities[2]. In most cases, the applicable laws are not specified to users.

The laws, however, do form the basis for moderation of the news comments, and the moderators are very familiar with them: hate speech, incitement and threats are commonly forbidden in news comments. In addition to hate speech targeted at ethnic and racial minorities, some countries' laws forbid insults and threats to groups or people representing a certain sex, sexual minorities, the disabled, religious groups or conviction status. Defamation, breaches of privacy, illegal links, pornography and other such contents are illegal, and therefore, they are not tolerated in news comments.

What is finally taken as actual hate speech depends on the local laws. In Finland, for example, the incitement to hatred law was altered in summer 2011, and it now also includes incitement spread through the Internet. The owner of the website can be held responsible for the content (Rikoslaki, 11:10, 511/2011). In practice, it is forbidden to 'spread knowledge, opinion or other messages, in which violence or discrimination of groups is deemed agreeable or desirable, or in which people are compared with animals, parasites etc., or in which sweeping statements are made of people being criminals, or inferior to others etc.' (Valtakunnansyyttäjänvirasto, Report 17/34/11, 11). 'Following this rule, the Finnish Iltalehti has post-moderated comments from their news

---

2   http://www.telegraaf.nl/reacties/huisregels/.

discussions, for example from a discussion concerning Finland-based youngsters of Somalian origin, who had committed a rape crime, the comment: 'I hope from the bottom of my heart that someone will execute those apes' (Anonymous comment, IL.fi 19.4.2012, removed from discussion).

These types of utterances were recalled in the interviews with moderators in all countries. They are considered illegal hate speech that will definitely be removed from the news comments. Yet, as an interview with a Dutch moderator illustrates, it is not always easy to decide where to draw the line between acceptable commentary and illegal hate speech, especially in news comments that relate to ongoing political debate:

> Well, there is this right-wing party in Holland, around Geert Wilders, and, he has some very strong opinions about certain aspects in society. Some of the articles around Mister Wilders we open for comments, these are the articles that give the moderators the most trouble. So, what if what Wilders is saying on the edge of what is legal, how should we treat what the commenters are saying about the same case. […] These are cases that are very difficult, if, always the case is on the edge of what is legal and is not, or what is appropriate or what is not, are the difficult ones. (De Telegraaf, 11/2012)

Hate speech remains problematic to moderators; at the same time for reasons of profitability, they are constrained by efficiency not to spend too much time deciding. The discussion needs to be kept going, so a moderator cannot take too long on one comment. Therefore, when seeing possibly illegal comments, moderators are more likely to remove them than to publish them (interview with Helsingin Sanomat, 05/2012). Moderators cannot know for certain whether a comment really is illegal or not, as final judgement can only be given by legal officers. However, they can make educated estimations.

Ethical codes for journalism represent the second type of regulations that steer moderation. These codes were especially mentioned in Finland and Sweden; in Sweden, they are a part of the training for moderators. The ethical codes state for example that the human dignity of a person, the under aged, and the offenders and victims of crimes need specific protection (Julkisen sanan neuvosto, 2014). For these reasons, in the aforementioned case, the Finnish Iltalehti was especially careful to remove comments concerning the youngsters of Somalian origin, who had been convicted of a rape crime with an under-aged victim (interview with Iltalehti, 10/2012). More generally, the moderators strongly believe that news media have to act responsibly in society by not allowing any cyberhate on their sites; thus, they remove content to protect vulnerable groups.

Finally the third type of regulation guiding moderation are the self-regulatory house rules. These rules are often even stricter than the laws or media ethical codes in order to avoid any possible problems. Therefore, various types of insulting, swearing, stigmatizing, assaulting, bullying, harassing, racist, sexist and indecent comments are generally prohibited. These house rules comply with the guidelines given to users; in this way, it is easier for moderators to explain why a certain comment was not allowed upon request. A British moderator explains this practice:

> Now I'm not saying they're all hate speech because hate speech has got a legal definition, but they're kind of in that area. You know, they're aggressive, they're bullying, they're rude, they break our comment rules, they'll have to come off. (Guardian, 1/2013)

According to the interviewed moderators, in the end, the editorial policy, quality standards and media brand play a significant role in what is tolerated in the comments and what is not, as the discussions form an inherent part of the news product. These standards can vary for different news media brands even within a single media company: One news site allows more commenting and more content than another, perhaps even allowing swearing, while another brand of the same media company would not tolerate it. The interviewed moderation company in the Netherlands regularly checks with various news organisations that its moderation is in line with the desired brand:

> Because, house rules, should be and most of the times are, very similar to the brand values of a company. So, you see that house rules of De Telegraaf are different from the house rule of Trouw. It would be very strange if it wasn't. (Novia Facts, 11/2012)

Media companies strive to support open discussions and free speech, but they also stress the importance of their house values and standards. These, next to financial and efficiency arguments, are ultimately the most important frameworks when designing and moderating the comment fields. Furthermore, moderators argue strongly for the legal responsibilities of the publisher and the publisher's right to decide what it publishes. It is commonly stated that since large-scale freedom of speech is guaranteed on the Internet, there is no need for news media to support people's unlimited freedom of speech, and in so doing, stimulate hate speech at their own expense. This viewpoint is summarised by the Swedish Dagens Nyheter:

> But for our part it is also a question of our brand value that certain things simply cannot be found

on our site. But definitely media also have responsibility in so far that we can take care that discussions and debates are of a good quality and are not racist or sexist or… We can never stop the discussions everywhere on the internet. But on the contrary, we can withdraw from helping to normalize that kind of speech through not allowing it on our site. (Dagens Nyheter, 10/2012)

The news media also stresses that it is not their duty to negotiate their rights as publishers with the users, who sometimes feel their comments were unfairly removed. As long as they are the owners of the product and invest in it, they only need to support the free speech rights of those who know how to behave, not the unruly one's rights.

The newspaper uses that freedom of speech. And that's also an inalienable right. As long as people write in Helsingin Sanomat and we make those publishing decisions, it is pretty straightforward and unambiguous. (Helsingin Sanomat, 10/2012)

I mean, we are investing in our news product, we are investing in our website, and it is a website that we try to exploit and we try to maintain. Basically it's our house, our rules. On the other hand, being a news organization, you also have some duties to the public. […] If you're free, you also have the responsibility to handle that freedom. So, there's free speech on Telegraaf.nl, but we have rules. (De Telegraaf, 11/2012)

Altogether, these statements indicate that comments that would possibly be acceptable according to the laws restricting hate speech and other forms of cyberhate, are often moderated and removed due to media's own guidelines, which complement official regulations. These self-regulatory guidelines are often necessary and beyond that they are preferred by moderators as solutions to prevent hate speech and cyberhate since laws do not and cannot settle all problems. The rights of the publisher to limit discussions on their news sites cannot be withdrawn from media either.

However, one might question whether discussion is being restricted too much. This question becomes pronounced when considering that comments sometimes are pre-moderated according to these self-regulatory guidelines, and particular news items are selected in advance for commenting based on the brand and quality goals of the news outlet. Such comment fields no longer enable free debate or deliberation in the sense of 'free speech' – instead, they normatively exclude the kind of discussants and content that could damage the news site's quality brand. These news discussions come close to the restricted public spheres of the civilized and responsible citizens that Habermas and Mill discuss, instead of being public spheres open for all debates and participants.

CONCLUSIONS AND DISCUSSION

This article has explored how various news media in Finland, Sweden, the Netherlands and the UK moderate cyberhate and hate speech. The study shows that moderation is not the only solution to prevent hate speech and cyberhate, but generally, the whole structure of the comment fields has been designed in a way that efficiently prevents problematic content. The number of available discussions and the practice of pre- or post-moderation dictate, how freely people can participate and comment. Finnish and British sites allow the most discussion, while Swedish and Dutch news sites prefer selected discussions, moderated by professional moderation companies.

The studied news media thus balance between preserving selected high quality discussions and attracting large amounts of users and comments. All selected solutions in a way try to guarantee advertisers' interest and revenues, but some are fairly weak in supporting free public debate. Common to all solutions is that the media is not willing to relinquish its traditional role as the gatekeeper of public debate, but wants to maintain in active control. Consequently, the public does not have an access to genuinely open public news debates online (see also Hermida &Thurman, 2007; Braun & Gillespie, 2011; Trygg, 2012).

This also indicates that media is willing to host a public discussion, but only according to certain normative rules. Like theoretical discussions have suggested earlier, there is a normative problem of selected participants and opinions to the liberal ideal of public debate, and even in the era of Web 2.0, the news media does not seem able to avoid the replication of the problem. This is regretful, considering that news comments have the potential of being a meeting point for a variety of people and opinions, a rarity on today's Internet.

In all countries, the actual practice of moderation is regulated by legislation on hate speech and related issues. Media ethical codes also guide moderation. Self-regulatory house rules complement these two practices and ultimately decide which comments are published. Self-regulation is preferred as a solution to cyberhate, since laws cannot settle all problems. Through self-regulation, many variants of cyberhate are removed from the discussions. Allowable content is defined by house rules, and they must conform to the brand and quality values of a news media outlet.

The studied news media are very aware of their role in preventing hate speech and cyberhate from being published; they see such practice as part of media's public responsibility. In this way, media is protecting vulnerable groups in society from being publicly offended. Media is following the general line of thought in Europe that recognises hate speech as an act of offense that can cause individuals and society serious harms (Calvert, 1997; Tsesis, 2002). For these reasons, moderation is justified.

However, if and when the media in addition moderates contents based on their quality guidelines, they enter into a problematic field. Moderation practices should be transparent and based only on laws and regulations that do not appear arbitrary (Benesch, 2013). This study has shown that there is a grey area in moderation and certain contents are removed just to be on the safe side. More research is still needed to observe the actual practice of moderation to find out, how considered the decisions to moderate are.

The news media are less keen to protect freedom of speech as an ultimate right in the comment fields. Media's own rights and the responsibilities as a publisher weight heavier than users' rights on free speech, since as it is argued, these can be practiced elsewhere on the Internet (see also Goodman, 2013). When defending such a view, the news media give away their initial ideal of news comment fields as fora for free expression of opinions. The ideals of free expression are in the end given less weight than the losses that would follow, if advertisers and users would abandon the media due to an inadequate quality of discussions (Hermida & Thurman, 2007; Braun & Gillespie, 2011; Trygg, 2012).

Allowing mainly qualitatively desirable discussions means that all opinions cannot be expressed freely in sites that traditionally form a part of democracy, namely the news outlets. Like previous research has demonstrated, this can on the one hand lead into discussions that are too clean on the outside and do not show what the actual societal concerns are. It can also lead into hidden racism, when new neat ways to express old ideals are developed (Hughey & Daniels, 2013). There is a risk that if certain opinions are not allowed in public debate, they will be expressed in more harmful ways elsewhere. Society needs public debate, also on sensitive issues, and the news media should still be able to provide a forum, where opinions can be expressed openly and safe by all.

MEJE SOVRAŽNEGA GOVORA IN SVOBODE GOVORA NA MODERIRANIH SPLETNIH STRANEH NA FINSKEM, ŠVEDSKEM, NIZOZEMSKEM IN V VELIKI BRITANIJI

Reeta PÖYHTÄRI
Univerza Tampere, Fakulteta za komunikacije, medije in gledališče, Raziskovalni center za novinarstvo, medije in komunikacijo (Comet), Kalevantie 4/E313, 33014, Finska
e-mail: reeta.poyhtari@uta.fi

POVZETEK

Članek primerjalno analizira strukturo in moderiranje komentarjev pod spletnimi novinarskimi prispevki na Finskem, Švedskem, Nizozemskem in v Veliki Britaniji. Osredotoča se na dilemo med svobodo govora in moderiranjem sovražnega govora in spletnega sovraštva. Študija temelji na intervjujih s 16 moderatorji in tekstovni analizi strukture in smernic moderacijskih praks 18 novičarskih spletnih strani. Novičarski mediji aktivno preprečujejo objavo sovražnega govora in spletnega sovraštva na različne načine glede na državo. Sovražni govor in spletno sovraštvo v komentarjih pod spletnimi novinarskimi prispevki sta podobna v vseh analiziranih državah. Moderacijska praksa temelji na treh vrstah pravil: zakoni, medijska etika in samoregulacijske smernice.

Ključne besed: komentarji pod spletnimi novinarskimi prispevki, svoboda govora, sovražni govor, spletno sovraštvo, moderacija.

INTERVIEW SOURCES (ANONYMOUS INTERVIEWEES):

Finland:

Helsingin Sanomat, Sanoma Oy, National newspaper. Interview (2 informants) 10/2012.

Ilkka, I-Mediat Oy, Local newspaper. Interview 10/2012.

Iltalehti, Alma Media Oy, Tabloid. Interview 10/2012.

Kainuun Sanomat, Alma Media Oy, Local newspaper. Interviews (2) 05/2012.

Uusi Suomi, Nikotiimi Oy, Online newspaper. Interview 10/2012.

Yleisradio (YLE), Public Broadcasting company. Interview 05/2012.

The Netherlands:

Novia Facts, Moderation company. Interview 11/2012.

De Telegraaf, TMG Landelijke Media B.V.,Tabloid. Interview 11/2012.

De Volkskrant/Algemeen Dagblad/Trouw, De Persgroep B.V., National newspapers. Interview 11/2012.

Sweden:

Aftonbladet, Aftonbladet Hierta AB, Tabloid. Interview 1/2013.

Dagens Nyheter, Dagen Nyheter AB, National newspaper. Interview 10/2012.

Sveriges radio, Public Broadcasting Company (radio). Interview 6/2012.

The United Kingdom:

Guardian, Guardian News and Media Limited, National newspaper. Interviews (2) 1/2013 and 2/2013.

REFERENCES

Back, L. (2002): Aryans Reading Adorno: Cyber-culture and twenty-first century racism. Ethnic and Racial Studies, 25, 4, 628–651.

Benesch, S. (2013): The Dangerous speech project. Dangerous speech: a proposal to prevent group violence. A paper presented at the 14th meeting of European Policy Planners' Network on Countering Polarisation and Radicalisation (PNN), Helsinki Finland, 25. 9. 2013.

Bleich, E. (2011): The freedom to be racist? How the United States and Europe struggle to preserve freedom and combat racism. Oxford, Oxford University Press.

Braun, J., Gillespie, T. (2011): Hosting the public discourse, hosting the public. When online news and social media converge. Journalism Practice, 5, 4, 383–398.

Calvert, C. (1997): Hate speech and its harms: a communication theory perspective. Journal of Communication, 17, 1, 4–16.

Cammaerts, B. (2008): Critiques on the participatory potentials of Web 2.0. Communication, Culture & Critique, 1, 4, 358–377.

Cammaerts, B. (2009): Radical pluralism and free speech in online public spaces. The case of North Belgian extreme right discourses. International Journal of Cultural Studies, 12, 6, 555–575.

Canter, L. (2013): The misconception of online comment threads. Content and control on local newspaper websites. Journalism Practice, 7, 5, 604–619.

Cooper, C. A. (2004): Cyber-hate and the disinhibiting effects of anti-gay speech on the internet. In: Linn, R. A. (ed.): Race/gender/media. Considering diversity across audiences, content and producers. Boston, Pearson Education, 258–265.

Council of Europe (1950): The European Convention of Human Rights. In: Council of Europe. Http://www.echr.coe.int/Documents/Convention_ENG.pdf (30. 5. 2014)

Council of Europe (1997): The Council of Europe's Committee of Ministers' Recommendation 97(20). In: Council of Europe. Http://www.coe.int/t/dghl/standard-setting/hrpolicy/other_committees/dh-lgbt_docs/CM_Rec%2897%2920_en.pdf (30. 5. 2014).

Daniels, J. (2009): Cloaked websites: Propaganda, cyber-racism and epistemology in the digital era. New Media & Society, 11, 5, 659–683.

Deuze M. (2003): The Web and its journalisms: considering the consequences of different types of news media online. New Media & Society 5, 2, 203–230.

Deuze M. (2006): Participation, remediation, bricolage: considering principal components of a digital culture. The Information Society, 22, 2, 63–75.

Deuze M. et al. (2007): Preparing for an age of participatory news. Journalism Practice, 1, 3, 322–338.

Domingo, D. et al., (2008): Participatory journalism practices in the media and beyond: an international comparative study of initiatives in online newspapers. Journalism Practice, 2, 3, 326–342.

Douglas, K. M. (2008): Psychology, discrimination and hate groups online. In: Joinson, A. N., McKenna, K., Postmes, T., Reips, U.-D. (eds.), The Oxford Handbook of Internet Psychology. Oxford, Oxford University Press, 155–164.

Downey, J., Fenton, N. (2003): New media, counter publicity and the public sphere. New Media & Society 5, 2, 185–202.

Edelstein, Y., Wolf, C. (ICCA) (2013): Report and recommendations of the Task Force on Internet Hate of the Inter-Parliamentary Coalitions for Combating Anti-Semitism (ICCA). In: Anti Defamation League. Http://www.adl.org/assets/pdf/press-center/ICCA-Report.pdf (30. 5. 2014).

Erjavec, K., Poler Kovačič, M. (2012): "You Don't Understand, This is a New War!" Mass Communication & Society, 15, 6, 899–920.

Fraser, N. (2007): Transnationalizing the public sphere. On the legitimacy and efficacy of public opinion in a Post-Westphalian world. Theory, Culture & Society 24, 4, 7–30.

Garland, J., Chakraborti, N. (2012): Divided by a common concept? Assessing the implications of diffe-

rent conceptualizations of hate crime in the European Union. European Journal of Criminology, 9, 38, 38–51.

Goodman, E. (WAN-IFRA) (2013): Online comment moderation: emerging best practices. In: Germany: Darmstadt, The World Association of Newspapers WAN-IFRA. Http://www.wan-ifra.org/reports/2013/10/04/online-comment-moderation-emerging-best-practices (17.9.2014).

Habermas, J. (1984 [1962]): The Structural transformation of the public sphere. Cambridge, MIT Press.

Hannula, I., Neuvonen, R. (2011): Internetin keskustelupalstan ylläpitäjän vastuu rasistisesta aineistosta(The responsibility of the forum owner for racist contents.). Lakimies, 3, 2011, 527–548.

Heinonen, A. (2011): The journalist's relationship with users: new dimensions to conventional roles. In: Singer, J. (ed.): Participatory journalism: guarding gates at online newspapers. Chichester, Wiley-Blackwell, 34–55.

Horsti, K. & Nikunen, K. (2013): The ethics of hospitality in changing journalism: A response to the rise of the anti-immigrant movement in Finnish media publicity. European journal of cultural studies, 16, 4, 489–504.

Hughey, M. W., Daniels, J. (2013): Racist comments at online news sites: a methodological dilemma for discourse analysis. Media, Culture & Society, 35, 3, 332–347.

Jankowski, N.W. & Wester, F. (1991): The qualitative tradition in social science inquiry: contributions to mass communication research. In: Jankowski, N.W. & Wester, F. (Eds.): A Handbook of Qualitative Methodologies for Mass Communication Research. London and New York, Routledge, 44–74.

Joinson, A. N. et al., (eds.) (2008): The Oxford Handbook of Internet Psychology. Oxford, Oxford University Press.

Julkisen sanan neuvosto (2014): Guidelines for Journalists and an annex. In: Julkisen sanan neuvosto. Http://www.jsn.fi/en/guidelines_for_journalists/ (30. 5. 2014). (The Council for Mass Media in Finland.)

Jørgensen, R. F. (2013): Freedom of expression in the Internet era. In: Carlsson, U. (ed.), Freedom of expression revisited. Citizenship and journalism in the digital era. Göteborg, Nordicom, 119–129.

Kortteinen, J. (1996): Sananvapaus ihmisoikeutena. In: Nordenstreng, K. (ed.): Sananvapaus. Juva, WSOY, 32–88. (Freedom of speech as a human right. In: Freedom of speech.)

Margolis, M., Moreno-Riaño, G. (2009): Democracy, tolerance and the internet. In: Margolis, M., Moreno-Riaño, G (Eds.): The prospect of internet democracy. Surrey & Burlington, Ashgate, 69–94.

Mill, J. S. (1982 [1859]): Vapaudesta. Helsinki, Otava. (On Liberty.)

Molnar, P. (2012): Free speech debate. Why hate speech should not be banned. Http://www.Freespeechdebate.com/en/discuss/why-hate-speech-should-not-be-banned/ (30. 5. 2013).

Mouffe, C. (2005): On the political. London & New York, Routledge.

Nieminen, H., Nordenstreng, K. (2012): Sääntely ja viestintäpolitiikka. In: Wiio, O. A., Nordenstreng, K. (Eds.): Suomen mediamaisema. Tampere, Vastapaino, 312–333. (Regulation and communication policy. In: Finnish media landscape.)

Nordenstreng, K. (2013): Deconstructing libertarian myths about press freedom. In: Carlsson, U. (ed.): Freedom of expression revisited. Citizenship and journalism in the digital era. Göteborg, Nordicom, 45–59.

Papacharissi, Z. (2002): The virtual sphere: internet as public sphere. New Media & Society 4, 1, 9–27.

Rikoslaki 1889:39, 11 luku (511/2011). In: Finlex. Finlex: Http://www.finlex.fi/fi/laki/ajantasa/1889/18890039001 (30.05.2014). (Finnish Criminal Law.)

Roversi, A. (2008): Hate on the net. Extremist sites, neo-fascism on-line, electronic jihad. Aldershot Hampshire, Ashgate.

Rønning, H. (2013): Freedom of expression is not a given right. In: Carlsson, Ulla (ed.): Freedom of expression revisited. Citizenship and journalism in the digital era. Göteborg, Nordicom, 13–25.

Sunstein, C. (2001): Republic.com. Princeton & Oxford, Princeton University Press.

De Telegraaf (2014): Huisregels. In: De Telegraaf. Http://www.telegraaf.nl/reacties/huisregels/. (Houserules.)

Torres, da Silva M. (2013): Online forums, audience participation and modes of political discussion: readers' comments on the Brazilian presidential election as a case study. Communication & Society/ Comunicación y Sociedad, 26, 4, 175–193.

Trygg, S. (2012): Is comment free? Ethical, editorial and political problems of moderating online news. Nordicom Information, 1, 34, 3–22.

Tsesis, A. (2002): Destructive messages: how hate speech paves the way for harmful social movements. New York, New York University Press.

Valtakunnansyyttäjänvirasto (2012): Rangaistavan vihapuheen levittäminen Internetissä. Rangaistavan vihapuheen määrittäminen ja rikosoikeudellisen vastuun kohdentuminen erilaisiin Internetissä toimiviin toimijoihin. Työryhmän raportti 21.12.2012. Dnro 17/34/11. In: Poliisi. Http://www.poliisi.fi/poliisi/helsinki/home.nsf/files/46428FC143119DE7C2257AEF00421F53/$file/17-34-11%20ty%C3%B6ryhm%C3%A4raportti%20finaali.pdf (30.05.2014). (The Office of the Prosecutor General: Spreading of penal hate speech on the Internet.)

Weber, A. (2009): Manual on hate speech. Council of Europe Publishing. In: Council of Europe. Http://www.coe.int/t/dghl/standardsetting/hrpolicy/publications/hate_speech_en.pdf (30.5.2014).

Young, I. M. (2000): Inclusions and democracy. Oxford, Oxford University Press.

Youngs, G. (2009): Blogging and globalization: the blurring of the public/private spheres. Aslib proceedings: New information perspectives, 61, 2, 127–138.