# “You had me there. Right up to the bit you were racist”:

## A Computer Mediated Discourse Analysis of Hate and Counter Speech in Social Media

Thesis Submitted for the degree of Doctor of Philosophy

*Luke Roach 2021*

*Cardiff University | School of Social Sciences*

# Content Warning

This thesis discusses sensitive topics, particularly racism and homophobia. Please be aware that extreme and offensive racial and homophobic slurs will be presented and discussed within this research. For the posterity of the data and its forensic analysis, examples of these hateful rhetorics are presented uncensored and in full as they were originally collected.

# Abstract

This thesis presents a computer mediated discourse analysis of online hate and counter speech found on social media. Social media and the ubiquitous connectivity of digital networks pervade day to day life in a way which allows near constant interaction and communication, and these technological developments coupled with world-wide socio-political transformations has led to a focus in popular debate and academic research on the production of hate speech and its governance. Of chief interest to this research is counter speech, a response to hate speech often hailed for its universality, its ease of use and its avoidance of potentially contentious censorship. While both kinds of speech are the focus of much academic research, there remains a gap in the qualitative space that focusses on the form and function of these speech acts - a gap which this thesis attends to. This thesis employs three distinct but complementary discourse analytic traditions to investigate hate and counter speech to understand how the aims of these kinds of speech are achieved in situ by online social media users. By looking at the use of identity, (im)politeness, and generic intertextuality in naturally occurring instances of hate and counter speech, this analysis finds specific recurrent discursive techniques that are deployed by speakers to achieve their goals. This thesis argues that in understanding their features and styles, hate and counter speech are identifiable as genres of speech in themselves. Users employ these genres online, and manipulate them through identity work, humour, politeness and impoliteness, in an attempt to produce effective hate and counter speech to achieve their communicative aims.

# Table of Contents

# Acknowledgements

First and foremost, I would like to thank my supervisors Professor Matthew Williams and Professor Pete Burnap. My research has taken some wild swings in methodological and theoretical focus over the years I've been doing it, and they have always provided me with great support, insight, and encouragement when I've needed it most.

Secondly, I would like to thank my progress reviewers. I've had a few over the years, and the discussions in each meeting always gave me something that vastly improved my research. I am extremely grateful for having had so many fresh and discerning eyes look over my work at its different stages.

I would also like to thank Dr Rob Smith who has given me a great deal of advice (and pints) over the course of my post graduate research, and whose guidance in struggling my way through a conference in New York managed to breathe new life into my thesis.

Huge thanks to my Mum and Dad and my brother Max. The unconditional belief you have all shown in me to finish this thesis has been hugely strengthening, especially in times when I wasn't so sure of myself. I'll never be able to repay what you all have done for me while I've been trudging through this thesis, but I want you to know it is always appreciated. Oh, and thanks for the constant financial support throughout my very extended university life. I guess keeping me from being destitute and homeless all this time was pretty nice of you too.

Thanks to my extended family too, my aunties and uncles, my bonkers cousins, my Nan and my grandad Ron. Ron taught me at a very early age that asking questions was cool, and that learning swear words was funny. I don't think I would have done any of this without those lessons.

Thank you to my girlfriend Marie, who has persistently motivated me and shown me an absolute wealth of patience and love while I've been locked away at my computer. You're a friggin' Angel.

Thank you to my housemates Jon and Amy. You've been invaluable help throughout the PhD, for academic reasons, for sitting on the sofa eating takeaway reasons, and for being drunk and strange around the world reasons. You two are smarter than me, more hardworking than me and younger than me, and I don't think I would have gotten through this without you.

Thanks to my home friends Gary, Matt, Si, Porkinson, Jamie and a load more. Cheers for showing a genuine interest in my work, no matter how many times I rambled a poor description of it at you over the years.

Thank you to Stav and Harriet. You two are the best, and Harriet, this is all your fault. You know how suggestible I am. Next time you bring something up at a party, don't make it something that's going to shape my life for the next 6+ years.

Thanks to Dan, for listening to my nonsense over the faint din of metal music and D&D soundscapes. Our musical and nerdy endeavours have been a much-needed distraction over the years, for which I am forever grateful.

And finally, thank you to my Cardiff friends, Eve, Anna, George, Bryony, Jason and Bryn. The range of ways you've all kept me sane and happy over the last few years would look crazy to list, but it has all meant the world to me. You're a darn good group. Darn good.

You're all darn good.

# CHAPTER 1: Introduction

As online sociality becomes ever more ubiquitous in day-to-day life, so too have the deviant aspects of social interaction that accompany it. Following popular worldwide developments not only in technology and its proliferation, but also in social and political polarisation around the world, the real and tangible impact of online deviance has been solidified as in important problem. A problem that necessitates a greater level of investigation and understanding to ensure the safety of users and its knock-on effects on wider issues offline. Not only is online deviance an increasingly prevalent and cognisant issue within society, but the advent of socio-political developments like Brexit and the rise of far-right populism in both Europe and the US has provided justification and legitimacy to fringe groups who proclaim isolationism, othering and hate speech in many forms.

The combination of the apparent socio-political climate of legitimised intolerance and the tools afforded by the internet provides a suitable venue for hate speech. The proliferation and normalisation of high-speed, long-distance communication through social media has amplified both the reach and volume of online hate, and those structures and systems require investigation. By taking racist and homophobic discourse as its site of primary inquiry, this thesis seeks to understand the impact that informal social controls, specifically counter speech, can have on combating hate speech. While both hate and counter speech have been under popular, academic and legal discussion for some decades now (Slayden and Whillock, 1995; Wasserman 2004), these forms of discourse are becoming ever more relevant for in depth, qualitative analysis, particularly with regards to their online iterations. Because of the scope and volume of possible hate speech online, much of the work done around its analysis and monitoring are focused on quantitative, big data studies (Williams and Burnap, 2015; Ayo et al., 2020; Laaksonen et al., 2020) that attempt to identify and track hate using automated and algorithmic techniques. Conversely, counter speech has not been subject to the same level of study, yet it is often touted as a useful and universal approach (Williams, 2019; Williams, 2021) to combatting hate speech. There remains a dearth of research (qualitative and quantitative) focussed on its definition, the understanding of its application in online settings and its effectiveness.

This thesis aims to provide a contribution to the fundamental qualitative work that must underpin the quantitative coding and counting work (Herring 2004:343) that proliferates much of hate and counter speech research. In engaging in a fine-grained computer mediated discourse analysis (Herring 1996, 2004) of instances of naturally occurring online hate and counter speech, this research identifies some of the key linguistic rhetorical moves used in situ by online citizens to achieve the aims of their discourses. Moreover, this thesis hopes to add to the understanding of the practical relevance of counter speech, providing useful and deployable examples of counter speech that can be utilised by interaction members online. By investigating and explaining the ways in which naturally occurring counter speech is performed, this work hopes to contribute to the developing guidance on counter speech protocols, to expand beyond the "appeals to reason" or "requests for evidence" (Williams, 2019:41) often proffered by counter speech advocates. In exploring the ways counter speech is already being designed by users in the wild, this thesis hopes to contribute towards understanding this form of hate speech mediation as an even more universally available technique.

## RESEARCH QUESTIONS

This research began with a broad research question, namely "How is hate speech and counter speech *done* by social media users?". By beginning with such a general focus, this research hoped to allow relevant phenomena within the data to emerge naturally, rather than its analysis being driven by more specific goals. Herring (2004) nods towards Glaser and Strauss' (1976) grounded theory approach noting that:

> "This approach is especially well suited to analysing new and as yet relatively undescribed forms of CMC, in that it allows the researcher to remain open to the possibility of discovering novel phenomena, rather than making the assumption in advance that certain categories of phenomena will be found" (p.358).

In mentioning grounded theory here, it is worth addressing the misattribution often assigned to grounded theory as discussed by Atkinson (2013). This research is not an entirely data-driven, a-theoretical endeavour, as the invocation of grounded theory could suggest. It is generated and informed by the data but brings with it the pre-conceptions and notions which drove the initial desire to engage with hate and counter speech data in the first place. And while the constituent parts of this research are not in themselves novel[1], detailed discourse analyses of online hate and counter speech are sparsely done, and it is a key epistemological stance of this study – to discover *how* these communicative actions are achieved based on what is found in naturally occurring data[2].

Herring (2004: 346) explicates what she believes are the 4 characteristics of a good CMDA research question: it should be empirically answerable from the data, non-trivial, motivated by a hypothesis and open ended. The overall research question stated above, and the subsequent specified questions detailed below, were each designed with these characteristics in mind. Each question focuses on what is observable as *being done* by interactants in the data. The analysis in each chapter is focused on what is observable in the utterances and interaction based on evidence within text. Furthermore, the research undertaken here is considered to be nontrivial and of use to further the understanding of how violent and combative speech occurs 'in the wild'. This is not a purely academic study, but one that hopefully has broader implications on, and provides the groundwork for, a better understanding of how to reduce harm. This deliberately expansive question ("how is hate speech and counter speech *done* by social media users?") was chosen to provide room for a data driven analysis, one which sought to discover what was occurring within the data and illuminate it without the constraints of a more specific focus from the outset. This thesis intends to argue that by engaging in fine detail with hate and counter-hate online 'talk', a greater understanding of the ways these discourses 'work', *in situ,* can be achieved.

[1] (Online Identity and politeness has been discussed by Sifianou and Bella 2019 and Garcés-Conejos Blitvich 2010, Online Intertextuality has been studied by Vasquez 2015 and Zidjaly 2010 etc.)

[2] Online communication, such as that discussed in this thesis, are not always considered to be "naturally occurring" as described here. Williams et al (2017) notes that the data gathered online can be understood as being manufactured to achieve certain aims and shaped by the architecture and policy of the platform they are found on. However, in this thesis they will be described as "naturally occurring" to denote that they were generated without prompting or input from the researcher or data collector.

The research question is also informed by different contextual factors. Firstly, a major imposition on the *doing* of hate and counter speech on social media are the limitations and key structural features of that medium. Social media platforms enforce their own terms of service and rules by which online citizens must conduct themselves, as well as technical restrictions such as character limits and sensitivity and quality filters. The rules and definitions of online platforms are expanded upon in the literature review (Chapter 2) upcoming in this thesis. While the data used for this research were collected in a time where language restrictions were less fervently enforced, there was none-the-less a degree to which certain kinds of speech were defined and managed on individual social media platforms. In the years leading up to the time period when these data were collected, Twitter implemented a series of measures attempting to curb hate speech, and police the quality of their posted content. In December of 2015 they officially banned 'hateful conduct' for the first time, and issued guidelines about what users cannot post under punishment of post deletion and/or account suspension. As a means to further stem the propagation of hate speech, Twitter introduced the 'quality filter' in 2016 and the 'hide sensitive content' feature in 2017. Both of these new features were set as default for all, in an attempt to hide hate speech from the majority of users and reduce the likelihood of retweeting 'unhealthy conversational content'. The broad research question stated above was therefore re-focused into the sub-question "How is hate speech and counter speech done by social media users, when faced with the limitations of social media platforms?"

In addition to the impositions and restrictions brought to bear by social media platforms, engagement with the data identified further limitations held against producers of hate and counter speech. The most pertinent to this research being the social stigma and popular consensus against hate speech (because of its detestable content) and to a lesser extent counter speech (because of its censorious nature and its connection to performative "wokeness" and political correctness). Counter speakers also found themselves up against the constraint of being a purely reactive endeavour, which must be generated and performed "on-the-fly" and in situ. Once again, the broad original question required reformulation to consider of how hate and counter speech are done under the social stigma of such discourses. There are a multitude of influencing factors that affect how a person engages in discourse and interaction, and those become more apparent and explicit when utilising a technology

that not only employs automated surveillance and censorship tools, but also open those discourses to the critical eyes of a much vaster audience than would be reachable offline. The task then became to try to discern in what ways discourses and linguistic strategies are invoked to achieve the communicative aims of their chosen forms of speech despite their unique restrictions.

With the broad understanding that the data dealt with in this research is, at the very least, under the weight of the technological and social influences described, a more focused set of research questions were developed. By taking a more grounded, inductive approach to discovering themes of inquiry, specific discourses within the data became apparent as useful and relevant in working towards answering the overarching research question. As Herring (2004: 358) suggests "one could let the phenomenon of interest emerge out of a sample of computer-mediated data". The themes that appeared to be most interesting and recurrent in the data were identity, politeness and generic intertextuality. In discovering these themes, more specific sub-questions were developed to help build an analytic picture which would aid in answering the original question. For clarity, when the term "generic" is used to describe hate or counter speech, that does not refer to 'cliché' or non-specific kinds of language, but speech which falls within the boundaries set by its *genre* definition. The full set of questions are as follows:

## Research questions

1. How are the communicative aims of hate speech achieved online, in spite of technological interception and social stigma?
2. How are the communicative aims of counter speech achieved online?
3. How is identity created, demonstrated and weaponised to achieve the communicative aims of hate speech and counter speech?
4. In what way are (a) politeness and (b) impoliteness manufactured and used as a tool for the production of hate and counter speech?
5. How is the intertextual gap between "generic" hate/counter speech and "naturally occurring" hate/counter speech manipulated?

The initial two questions address the broad analytic focus of this thesis, discerning how hate speech and counter speech are done by social media users online. The first question relating to hate speech was constructed with additional specification when compared to its counter speech counterpart to acknowledge the additional situational influences imposed on hate speech as a policeable and socially maligned form of speech. It is worth noting that counter speech can itself be produced in a hateful way and may fall under these constraints. However, as will be discussed in the upcoming literature review, by its working definitions counter speech is not assumed to be, or proscribed to be, an innately hateful form of speech. It is for this reason that in the framing of these research questions hate speech is assumed to be universally policeable by technology and social pressure, while counter speech is not.

Question 3 was decided upon after making the arguably obvious but important observation that the creation and co-creation of identities was a central and frequent occurrence within the data. When addressing issues of hate speech, identity, othering and the classification of at-risk groups will always play a major part in its performance. However, what became apparent during engagement with the data was the frequent invocation of identities constructed in situ, and their use as a tool to achieve the communicative aims of their discourses. Interactants were invoking different transitory identities, for themselves, for their targets, for their audiences and for their combatants which aided in the execution of their chosen form of speech. Beyond the common categories expected when discussing hate speech (racial, national, sexual and gender categories to name but a few), hate and counter speakers were found to invoke identities which provide authenticity for the self and weaponize delegitimisation against the other.

Question 4 was again designed after extended engagement with the data illuminated the interesting dichotomy of seemingly oxymoronic occurrences of politeness in hate speech and impoliteness in counter speech. Again, politeness and impoliteness in discourse is a well-worn and not particularly novel area of study, but what proved interesting was that in comparison to much inquiry into politeness strategies, hate speakers found in the data were engaging in linguistic politeness while trying to actively ‘face damage’. Classically, politeness is used to minimise damage in an unavoidable face threatening situation, whereas in this instance the face damage is the entire aim of the speech act. Contrastingly, in observing the instances of

counter speaking, a speech act whose main focus is to oppose othering and exclusionary speech, many examples were found in which speakers engaged openly and intentionally with face threatening impoliteness. Once again, the contrast between communicative aim and linguistic execution appeared novel and worth inquiry.

Finally, question 5 is informed by the tension between the generic forms of each kind of speech (i.e. hate or counter speech as exemplified and constrained by their genre definitions) and the ways in which social media users materially achieve the communicative aims of their speech under the pressure of social stigma and online policing. By viewing hate and counter speech as the Swalean (1990) conceptualisation of genres of speech, it became apparent that the definitions used to describe and police those kinds of speech, while useful, were not an exhaustive and complete encapsulation of how those discourses naturally occurred. Speakers engaging in these speech genres were able to, consciously or not, manipulate the intertextual gap (Hodges, 2015) between their performed speech and the prescribed definitions of those genres to achieve their aims in a variety of ways. Stubbs (2015) provides a concise summation of this idea when he notes that

> "…we find a text easy to understand if it consists of familiar topics being talked about in ways that are familiar from other texts. In a word, understanding depends on both our textual and our intertextual competence…. Conversely, we find a text difficult to understand if it is lexically and semantically dense: that is, if there is too little repetition of vocabulary, and if too many of the words are unfamiliar or are used in unusual combinations" (p.486)

It seemed interesting then that when users were engaging in a form of speech that would be familiar to others as inflammatory or policeable they would make use of the intertextual gap between the expected and the unexpected to achieve the aims of their speech at a metaphorical distance. Similarly, counter speech is often described and defined as directly challenging and in need of intellectual argument as a pre-requisite. As a form of online social regulation lauded for its universality, it was again noted how much speech (within the data) understood to be "counter" occurred at an intertextual gap to its generic definitions. With

the questions driving this research explained and justified, this chapter turns finally to a brief explanation of its structure.

## HOW TO READ THIS THESIS

Following this introductory chapter, this thesis moves on to an extensive overview of the literature (chapter 2) that has informed and shaped this research. This chapter delves into the different sociological, criminological and linguistic work that has been brought together to illuminate the different facets that amount to a study of online hate and counter speech.

Beginning with a synthesis of the many disparate characterisations of both hate and counter speech, this chapter solidifies a working definition of each discourse, on top of which the analysis is built. From there, the discussion progresses onto the topic of digital criminology, describing classical and updated criminological theory in terms of the online society this study finds itself in. Following on, this chapter investigates the key defining features of online social media which shape interaction and mark it as different and uniquely fascinating for this kind of academic research. Finally, this chapter enters a discussion of the theory informing each of the specific analytic lenses taken by this research. Looking at theories of Identity and Shame, Humour, Genre and Intertextuality, and linguistic Politeness and Impoliteness, this section illuminates the ground-breaking work done to develop and solidify these theories as robust and informative analytic frames. In particular, this section seeks to justify the classification of hate and counter speech as genres in themselves, and problematise the classical understanding of genre theory to provide for a more constructionist approach.

From there, this thesis advances onto chapter 3, a discussion of the methodology undertaken in this analysis. This chapter explicates the choice to embrace Computer Mediated Discourse Analysis as its main methodological approach, and the specifics of Symbolic Interactionism as a guiding principle by which these methods are conducted. This chapter also covers an in-depth discussion of the pragmatics of the data collection and presentation, addressing both the procedures and the justification of those procedures in

terms of the chosen methodological theory. From there this chapter addresses the individual methods employed for each of the analytic traditions applied to the data, before finally moving to a discussion of the current ethical landscape of online social media research and the ethical standards enforced for this research in particular.

Following on from the theoretical and methodological grounding of these chapters, this thesis turns to the analysis proper. The analysis in this thesis is split into two distinct findings parts, i) hate speech and ii) counter speech. Beginning with hate speech, this part is made up of three chapters covering each of the analytic traditions applied to the data; Chapter 4: Genre and intertextuality, Chapter 5: Identity and Chapter 6: (Im)politeness. Each of these chapters presents and analyses the data, using their particular lens to interpret the different linguistic and rhetorical techniques used by hate speakers in the wild. Chapters 7,8 and 9 constitute the counter speech part of the findings and follow a similar structure as the previous, presenting data and applying each of those analytic traditions, to discover how counter speech is achieved in situ. Each of these chapters and their key themes within each section can be visualised as follows:

- Hate Speech
    - Chapter 4: Genre and Intertextuality
        - Legitimation
        - Excuse Making
        - Projection
    - Chapter 5: Membership Categorisation Analysis and Identity
        - Declarative Identity
        - Oppositional Identity
        - Weaponising Identity
    - Chapter 6: Politeness
        - Markers of Solidarity
        - Going “Off Record”
        - Redressive Action

Fig. 1

- Counter Speech
    - Chapter 7: Genre and Intertextuality
        - Joke/Comedy
        - Mimicry
        - Reductio Ad Absurdum
    - Chapter 8: Membership Categorisation Analysis and Identity
        - Declarative Identity
        - Oppositional Identity
        - Weaponising Identity
    - Chapter 9: Impoliteness
        - Sarcasm/Mock Politeness
        - Positive Impoliteness
        - Impolite Beliefs

Fig. 2

Finally, Chapter 10 draws conclusions from the analysis, identifying the recurrent themes found throughout the data and using them to address the questions mentioned in the outset. This chapter also discusses the limitations encountered in this study, as well as make suggestions for future work that may build on or compliment what was discovered here. With the aims and questions driving this study explained, this thesis now turns its attention to a discussion of the literature which informed and supported this research.

# CHAPTER 2: Literature Review

## INTRODUCTION

Online social research is, as are all aspects of social science research, a constantly evolving and multifaceted realm of enquiry. Because of the exponential development of networked technology, academic research and legal policy have been striving to attend to the sociological and criminological implications that have arisen. Internet technology has fundamentally changed a great deal how we interact with one another, and through its networked ecosystem has provided a whole new world in which deviance and crime can occur. This development has not only altered how crimes are committed, but what kinds of crime can exist. Focusing in on the site of inquiry for this thesis, social media has allowed a more connected, unfettered flow of communication between disparate groups across the world, and with that the possibility and tendency for more aggressive and deviant communication has seemingly expanded too. To properly lay the groundwork for the upcoming analysis into hate and counter speech, this chapter is split into four sections, each discussing the relevant literature around the academic developments of hate and counter speech and digital social interaction.

This chapter begins with a thorough discussion of the base concepts surrounding hate and counter speech. Particularly, this section contends with the issue of defining these different kinds of speech, taking cues from the legal world, definitions from social media platforms themselves and drawing on key academic work from Williams (2019, 2021), Chakraborti and Garland (2012), Garland and Chakraborti (2012) and Richards and Calvert (2000). The fundamental understanding of what this thesis will take to mean by the terms “hate speech” and “counter speech” will provide a firm grounding from which one can comprehend the criminological backdrop of the online world and the specific analytic theory that drives this research.

The second section interrogates the developing criminologies which proliferate and most acutely align with this aspect of digital society. Chiefly among the theorists considered this section, Garland's ideas of the Culture of control (2001) and "responsibilization" will be used as a key platform to understand how the criminogenic aspects of digital life have developed. The scope and unique spatial temporal aspects for online social life have provided challenges for both the pro-active and re-active elements of police response to crime, and updates on classicist, every day criminologies like routine activities (Cohen and Felson, 1979) and rational choice (Cornish and Clarke, 1987) provide a useful base from which to understand the choice for pro-active deterrence favoured by many online platforms.

Following on, this chapter turns to a discussion of those online platforms and the key features of the design of social media. This section will be driven primarily by a discussion of Kiesler et al. (1984) social-psychological description of distance, anonymity and a-synchronicity in online communication when compared to offline interaction.

This chapter then considers the literature foregrounding the analytic traditions driving this research, investigating the theoretical underpinnings that drive these forms of discourse analysis while discussing their application and usefulness in an online setting, considering shame and identity. Next, this chapter will turn to a discussion of the Swalean conception of "genre" (1990), unpacking his criteria designating a style of language as a genre and demonstrating how hate and counter speech fall within those boundaries.

Following on, the literature of 'humour', a key theme found within these genres of speech, is outlined to understand the rhetoric style that can be used by both hate and counter speakers in attempting to achieve their goals. Finally, this chapter will end with a discussion of linguistic politeness and impoliteness, driven primarily by the work of Brown and Levinson (1978) and Culpeper (2005), before concluding the chapter.

## HATE CRIME AND HATE SPEECH

Hate crime, despite being a widely discussed concept used frequently and casually in public and popular dialogue, is still subject to academic debate about its definition and its conception as an identifiable phenomenon (Chakraborti and Garland 2012, Garland and Chakraborti 2012). Chakraborti and Garland suggest that the core defining feature of hate crime is the identification and victimisation of groups perceived to be subordinate to the aggressor. Individuals identified as fitting with subordinated or subjugated subcultural groups like the disabled, members of the LGBTQ+ community and those of minority ethnic status, can be understood as being at risk of hate crime victimisation. Garland and Hodkinson (2014: 3) note how scholars have considered an expansion of the parameters of hate crime victim types to include "the homeless, the elderly and those from alternative subcultural communities", groups who are not currently covered by hate crime law, but may still be recorded by police. The emphasis in identifying at-risk groups is on vulnerability and variance from the dominant social group, rather than the specific qualities of members within the group. Hate crime can be understood in terms of the imbalanced social relationships between perpetrators of crime and violence and the groups at risk of victimisation. In the legal practice of defining the victim perpetrator relationship, there are two discrete models that are generally accepted globally, but each is oriented to understand the offence differently. The "hatred motivation model" (Walters et al., 2017: 201) prescribes 'animus' or hostility towards a person based on their identifying characterises as necessary for designation as a hate crime, while the "group selection model" (Walters et al., 2017: 202) requires only that the victim is selected because of their group membership. The group selection model 'net-widens' to include more hate crime cases, since there is no need for explicit evidence of identity-based hostility, only that they are selected because of their characteristics.

It is important for this thesis to understand the current definition adopted by criminal justice institutions through which they intervene and analyse ongoing hate crime. Taking a victim focus, The National Police Chief's Council (NPCC) define hate crime as "Any hate incident, which constitutes a criminal offence, perceived by the victim or any other person, as being motivated by prejudice or hate" (Hate crime | The Crown Protection Service. 2021).

The Public Order Act (1986) also define incitement to racial hatred when they state that it is an offence when:

> "A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if— (a)he intends thereby to stir up racial hatred, or (b)having regard to all the circumstances racial hatred is likely to be stirred up thereby."

These definitions do not require confirmation of hateful intent[3] from the offender and make no claim or definition of evidence that a crime is motivated by hate, but instead require only that the crime is perceived to be hateful either by the victim or a bystander. Additionally, this definition makes no explicit reference to the power imbalance and requirement for the victim to be part of a specific group for this to be defined as a hate crime, only that the victim perceives prejudice to be a motivating factor. Consequently, definitions of hate crime, encompassing racial and religious hate crime were specified in the Crime and Disorder Act 1998, while it was not until the Criminal Justice Act 2003 that hate crimes motivated by disability, sexual orientation or transgender identity were officially defined[4]. Hate crime investigation currently holds a priority in UK policing which has spawned the creation of hate-specific policy such as the Hate Crime Action Plan (2012) in England and Wales and the Hate Crime Framework for Action (2014) in Wales

Race (Phillips and Bowling, 2012) and religious (Chakraborti and Zempi, 2012) hate crime has been the dominant field of study for most of the academic corpus, however this body of literature is expanding to include more examination of homophobic hate crime, as well as gender, elder, disability and cyber hate, as they become more widely reported and recognised. The Crime Survey for England and Wales (CSEW) 2019/20 reported that there were 105,090 incidents of hate crime recorded by the police in England and Wales, an 8%

[3] The Crown Prosecution Service's (2020) legal guidance on the prosecution of hate crime state that "where a demonstration of hostility can be proved, there is no need also to prove the motivation", but also suggest that the demonstration of hostility may or may not be the conduct element of the offence.

[4] Additionally, the Anti-Terrorism, Crime and Security Act (2001) expanded legal definitions to include racially or religiously aggravated assault, criminal damage, public order offences and harassment, the Criminal Justice Act (2003) enacted sentence enhancements where crimes were aggravated by hostility towards race, religion, sexual orientation or disability, and the Legal Aid, Sentencing and Punishing Offenders Act (2012) expanded to include hostility towards transgender identity.

increase from 2018/19. Combining the reports from 2017/18 and 2019/20, the CSEW estimates 190,000 incidents of hate crime a year, representing around 3% of all CSEW crime which itself is around 6.1 million incidents. They also note that hate crime reporting rates have been consistently higher (47%) than for all CSEW crime (38%). A possible justification for the prioritising of hate crime investigation may be because studies have pointed to an increased impact felt by victims of hate crime. Corcoran et al. (Home Office, 2015) suggests that hate crime is more likely to incur an emotional effect on the victim than other forms of crime and House et al. (2011) found that hate crime victimisation can lead to an increase in the frequency of suicidal thoughts. The CSEW 2019/20 also found a high level of repeat victimisation, with 27% of household hate crime victims being subject to repeat attacks. They also found a low level of satisfaction with the police, with 55% of hate crime victims feeling satisfied compared to 66% of crime victims overall, and 27% reporting that they were 'very dissatisfied' compared to 17% overall. These additional impacts may be, as Levin and McDevitt (1993) discovered, due to the hate crime frequently being found to include serial or repeat attacks from a multitude of offenders as opposed to discrete and singular victimisation. Wilkinson (2001) and Cardozo et al. (2003) also found that hate crime victimisation increased the likelihood of retaliatory and retributive action from victims, creating a violent offending loop that is less likely to occur in other forms of crime.

When discussing cyber-hate, the most common form of such interaction is production and dissemination of hate speech through social media platforms. McDevitt et al. (2002) and Levin (2002) found that the ubiquity of online media, as well as its lack of regulations (at least in the US) and necessary resources made it an easy and accessible tool through which to perform these kinds of crime. Perry and Olsson (2009) found that because online communities shared ideologies, interests and values rather than the traditional terrestrial constraints of family or geography, cohesive hate groups could be gathered from wide reaching geographical distances and solidified online into legitimate subcultures. Tell MAMA (2019) reports that 327 of their 1072 verified cases of Islamophobic hate speech occurred online. This is a drop of 10% from their figures in 2017, which coincides with an 11% reduction in reports of offline abuse. They theorise that this may be due to the 4 major UK terrorist attacks that had occurred in 2017, each of which brought spikes of hate crime reporting. When compared with their 2018/19 report, Stop Hate UK (2019/20) have reported a 17%

increase in race incident reports, 37% increase in homophobic incident reports and 600% increase in misogyny incident reports. In their survey of 1166 LGBT+ people, GALOP (Hubbard, 2021) found that 64% of respondents had experienced anti-LGBT+ violence or abuse. The most common forms of abuse reported were verbal abuse (92%) followed by online abuse (60%).

Hate speech and counter speech, as a specific focus within hate crime and its policing, require their own definitions so as to be readily identified, understood and intervened upon. The development of these definitions has had input from a plurality of sources, particularly because of hate crime's proliferation online, where policing structures are themselves more multi-agency. This chapter now moves onto a thorough overview of hate speech definitions, drawing on official legal definitions, as well as those produced by social media platforms to build an understanding of what these institutions believe constitutes hate speech, as well as how those definitions can be investigated by researchers and manipulated by speakers online.

## Defining Hate Speech

Williams's (2019) report into online hate speech conveniently collected a large majority of the working definitions. Drawing on the UK Crown Prosecution Service (CPS), European Council, and social media platform definitions, the report outlines the differences between what is deemed criminal hate speech, and that which does not reach that bar for legal intervention but is still considered governable by social media platforms. Recommendation (97)20 of the Council of Europe (1997) state that:

> "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" (p.107)

The Additional Protocol to the Convention on Cybercrime (Council of Europe, 2003) also state that:

> "racist and xenophobic material" means any written material, any image or any other representation of ideasor theories, which advocates, promotes or incites hatred, discrimination or violence, against any individual or group of individuals, based on race, colour, descent or national or ethnic origin, as well as religion if used as a pretext for any of these factors." (p.2)
>
> "distributing, or otherwise making available, racist and xenophobic material to the public through a computer system." (p.2)
>
> "threatening, througha computer system, with the commission of a serious criminal offence as defined under its domestic law, (i) persons for the reason that they belong to a group, distinguished by race, colour, descent or national or ethnic origin, as well as religion, if used as a pretext for any of these factors, or (ii) a group of persons which is distinguished by any of these characteristics" (p.3)
>
> "distributing or otherwise making available, through a computer system to the public, material which denies, grossly minimises, approves or justifies acts constituting genocide or crimes against humanity" (p.3)

This research's analytic focus then will be centred around speech which seeks to justify itself and attempt to avoid identification under these definitions while achieving the communicative aims of hate.

The UK Crown Prosecution Service identifies the difficulty of creating a solid and universal definition of criminal hate speech. Instead of giving a specific framework of language or communicative purpose, they instead provide a list of factors which must be taken into account when a prosecution is being considered. These include speech which is:

> “motivated by any form of discrimination or hostility against the victim's ethnic or national origin, gender, disability, age, religion or belief, sexual orientation or gender identity” (Williams 2019:14).

Hate crime covers offences that are aggravated by reason of hostility towards the victim based on race, religion, disability, sexual orientation or transgender identity. The range of legislation for prosecuting hate crimes includes the Crime and Disorder Act (1998), the Criminal Justice Act (2003), the Malicious Communications Act (1988) and the Protection from Harassment Act (1997). When social media posts don’t amount to specific offences in their own right (threats to kill, blackmail, harassment etc), they will still be considered criminal if their content is grossly offensive, threatening, abusive or intended to stir up hatred based on race, religion or sexual orientation. Hate speech prosecutions operate at a high evidential threshold, and consider whether prosecution is in the public interest based on the nature of communication and the impact of the victim.

Williams states that legal academics, such as Greenawalt developed a group of “expressive criteria” which can identify hate speech as criminal. In his paper Free Speech Justifications (1989: 122), Greenawalt notes that to keep speech ‘free’ “the minimal principle of liberty establishes that the government should not interfere with communication that has no potential for harm”. It follows then that the criteria for criminalising a particular kind of speech focuses on that which can justifiably be seen to do harm of some kind. The criteria listed state that for it to be deemed criminally harmful it must:

- Deeply wound those targeted
- Causes gross offence to those that hear it, beyond those targeted
- Has a degrading effect on social relationships within any one community
- Provokes a violent response

The definition of harm here is layered, requiring individual wounding, intra-community offence and inter-community degradation. These criteria also suggest the presence of a provocation to violent response. Definitions of violent response will be further explored in a

later section on counter speech, but at this point it is worth noting something which will be expanded upon in the discussion around genre and intertextuality. The assignment of language generic definition does not sit merely in the text and its intention but is a co-constructed phenomenon that is influenced and informed by its occasioned response. The provocation of a violent response, to whatever degree, is instrumental by these criteria for the designation of language as criminally hateful.

Continuing on, the Council of Europe (1997) state the following in definition of hate speech:

> "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." (p.107)

In this definition, "discrimination or hostility" which was found above is expanded to any expression which "spread, incite, promote or justify" hatred towards specific groups. This expands what counts as hate speech from that which directly attacks, to that which can be seen to promote or justify an attack. Possible arguments suggest that a hateful attack can be recognised as propagating the validity of an attack to those beyond the targeted individual or group, reinforcing again what was determined in the CPS definition. Additionally, this definition expands from a one-way damaging attack, to include speech which subjugates an individual or group by promoting supremacy of another more dominant group (as seen through "aggressive nationalism and ethnocentrism"). This definition also focuses solely on racial and xenophobic hate speech, failing to include other target groups mentioned above.

Combining these definitions, one can suggest that a generic ideal of hate speech is one that focuses on harm caused to a group identified by a vulnerable status within generally accepted social categories. That is to say that race is an accepted social category (as assumed in the legal definitions discussed above), and black and minority ethnic people within that category are susceptible to increased vulnerability in the UK, USA and western Europe,

because of historical, cultural and socio-economic factors. Similarly, sexual orientation and gender are seen as accepted social categories, and those who identify with LGBTQIA+ status may also experience increased vulnerability due to historical and current social prejudices. This combination also suggests that to fall within the genre of hate speech, one must identify harm as caused by direct attack or by indirect harm caused by the degradation of one group at the expense of the promotion of another. Williams goes on to note that in the UK speech must be "grossly offensive and/or inciting others to hate in a threatening or abusive way" (2019:15) to be considered illegal. Specifically, racial hate speech needs to be threatening *or* abusive, while other forms of hate speech need to reach the higher standard of being both threatening *and* abusive. Both criteria require a response from an audience or target to be defined as hate speech, once again speaking to the co-constructed nature of this genre.

### Social Media Definitions

Many social media platforms have introduced their own working definitions of hate speech with which to monitor the content on their sites. Most mainstream social media platforms have their own definitions, with Reddit being one notable exception, boasting no centralised definition, but deputising "subreddit" communities within the platform to decide and enforce their own terms for policed speech. Alternative or "fringe social media sites, such as Gab, Voat and 4chan" (Williams, 2019:16) identify themselves specifically as platforms without centralised speech policing mechanisms and bastions of free speech (Gab Social. 2020).

Beginning with YouTube, their definition is the shortest, with the least specificity but comes with a list of examples to illustrate what they find to be archetypal hateful content. Their working definition (hate speech policy - YouTube Help. 2020) states that:

> "We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of major violent events and their kin, Veteran Status"

This definition specifically identifies the "promotion" of hatred and violence, suggesting an inclusion of direct attacks and the propagation of hateful language. The inclusion of the un-qualified term "violence" in a definition of hate speech implies a broader definition of violence than specific physical harm that is often considered to be the standard, linking to the ideas of illocutionary speech as violence as discussed by Butler (1997) among others. This definition also mirrors that provided by the CPS, in that it considers hate speech to be uni-directional attacking speech, and not inclusive of speech that denigrates as an effect of aggrandizing a dominant category group. The attribution list, while again not identifying the groups within the categories that are deemed to be vulnerable, is quite expansive, even going as far as to include those with "veteran status" to be worthy of protection from harmful speech as well as speech that targets "victims of a major violent event and their kin". Although the guidance does provide 15 generic examples of hate speech, only 2 are provided in the main body of their definition, and they are:

> "I'm glad this [violent event] happened. They got what they deserved [referring to persons with the attributes noted above]."
>
> "[Person with attributes noted above] are dogs" or "[person with attributes noted above] are like animals."

The prominence of these two examples does not necessarily imply a priority in opposing these specific kinds of hate speech, but it does indicate that they believe them to be worth viewing when explaining hate speech to its users. Specific examples that could fit these generic moulds include praising the holocaust in terms of violent events aimed at a specific group of people, or the comparison of immigrants to swarming insects in the dehumanizing second example.

Both Facebook and Twitter report much more in-depth explanations of what they consider to be unacceptable hate speech, complete with definitions, rationales for their policing of these forms of speech, caveats where possible confusion or exceptions may occur, and examples of different forms hate speech may take. These discussions of the different forms of hate speech cover not only the differing content that can be identified as hateful

(threats of violence, use of slurs and offensive language) but also different mediums through which hateful content may be displayed (text, image, video, profile features).

Twitter (hateful conduct policy. 2020) describes "hateful conduct" as anything which would:

> "promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."

Twitter's guidance expands upon these categories in their "Rationale" section, to state that the protected groups within those categories are disproportionately targeted for hateful abuse, and that those who identify with any or multiple underrepresented groups may experience a "higher impact" from the abuse they suffer. The protected groups they identify include "women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities". The conduct policy then expands to describe the sorts of language which fall under their above definition, identifying "violent threats", "wishes or calls for serious harm to a person or group", "references to mass murder or violent events where protected groups have been the target", "inciting fear about a protected category", "repeated and/or non-consensual slurs, epithets, racist and sexist tropes or other content that degrades" and "hateful imagery". Each of these hate speech forms comes with an explanation and in most cases some examples.

Again, these examples all centre around the focal tenant of perpetrating harm against a vulnerable group, and interestingly this definition once again focuses on uni-directional harm that is an active attack and does not explicitly include speech which denigrates one or more groups through the promoting of a dominant one. Another feature of note concerns "inciting fear", where it is the fearful response from a category group not being attacked that defines the language being used to describe a protected category as hateful. In examples of this kind of text, although they contain negative valued judgements about a protected group, what is definitionally most important here is that those descriptions convince (or at least attempt to convince) a presumed audience to hate and fear a particular group. This is also the

only definition and explanation of hate speech that explicitly identifies “slurs and epithets” as kinds of hate speech, with all others employing broader definitions referring to dehumanising, mocking or comparative kinds of speech.

Following the defining and exemplifying sections, the guidance sets out the caveats for its application of these definitions, noting that the use of certain slurs, dependant on context, may not be identifiable as abuse but instead instances of the reclamation of historically targeted language by the vulnerable group (Ritchie, 2017). The acknowledgement of this kind of exception reinforces what was mentioned in the discussion of genre as a classifying tool; that the specific form, structure and text are not enough to adequately identify a genre, but an analyst must also engage with the context and the communicative purpose of those forms.

Finally, Facebook provide their own succinct definition of hate speech (Community Standards | Facebook. 2020) which also includes alongside it an explanation of how they define their use of the word “attack”:

> “We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability. We protect against attacks on the basis of age when age is paired with another protected characteristic, and also provide certain protections for immigration status. We define "attack" as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation.”

This definition once again identifies the general categories by which victims are identified, not the vulnerable groups within them. The proceeding definition of attack invokes notions of injurious speech that were more subtly nodded towards in previous social media definitions. “Violent or dehumanising speech” refers directly to the notion that words in themselves have direct consequences and are not merely symbols representing other ‘proper’ violence.

Facebook also presents a range of caveats to their identification of specific language as hate speech, expanding on Twitter's policy to include shared hateful content to raise awareness or educate, self-reference or empowerment (slur appropriation), gender exclusive language for control of group membership, humour and social commentary. Addressing slur appropriation and contextual caveats for the use of certain language, Facebook state that "we believe that people are more responsible when they share this kind of commentary using their authentic identity", suggesting that authenticity and visible identity result in more reasonable use of possibly policed speech. This is then followed by a detailed breakdown of hate speech into three tiers of severity, each containing examples. A slightly amended version of Williams (2019:16) summary is provided:

Tier 1 hate speech is that which:

- Is violent
- Is dehumanising and makes comparison to filth, bacteria, disease or faeces; to animals; to sub-humanity
- Mocks the concept, events or victims of hate crimes

Tier 2 hate speech is that which:

- Implies physical deficiencies related to hygiene or appearance
- Implies the inferiority of a person's or a group's mental or moral deficiency
- Expresses contempt or disgust
- Uses slurs [cursing or profanity] directed at a person or group of people who share protected characteristics

Tier 3 hate speech is that which:

- Calls to exclude or segregate a person or group of people who share protected characteristics that do not fall under general criticism of immigration policies and arguments for restricting those policies
- Describes or negatively targets people with slurs, where slurs are defined as words commonly used as insulting labels for the above-listed characteristics

A slight amendment was made to distinguish the tier 2 use of slurs as described as cursing or profanity in the official guidance, as distinct from the use as derogatory nouns associated with a protected group in tier 3.

With each of these definitions understood, and with reference to the examples provided within the expanded guidance, one can draw out some consistent features which are indicative of an utterance or piece of text's categorisation as hate speech:

Identification: hate speech commonly identifies a vulnerable group
Description: hate speech commonly uses negative descriptive or comparative language
Injury: hate speech commonly deploys violent, threatening and attacking language
Incitement: hate speech commonly incites others to share judgement of a group

These four features, based on the definitions above, are sought out when deciding if an utterance can be generically classified as hate speech. Hate speech identifies its target individual or group, negatively describes that group using comparison or stereotype, injures the target with violent and aggressive language, and promotes the same judgement and violence to others who may be audience to the interaction. While this is not an exhaustive list of all the communicative acts that are done or can be done during hate speech, these are suggested to be the most common and most archetypical of hate speech use. As will be shown in the data, a great deal of additional linguistic moves are used to accomplish hate speech and to engage in the manipulation of intertextual gaps around this generic framework. With hate speech definitions discussed and their application both legally and by online platform understood, this chapter will now turn its focus towards investigating the definitions of counter speech.

## Defining Counter Speech

Counter speech, in comparison to hate speech, is less tightly defined. Counter speech definitions tend to focus more on what the speech does and what it achieves than the specific vocabulary it is comprised of. While certain kinds of language that can be used as counter

speech may be contentious or policeable (for example when a different kind of hateful language is used to form an argument), counter speech as a concept or as a linguistic strategy does not experience the same legal or social stigma, and is therefore arguably not subject to the same level of scrutiny. While being the recipient of counter speech may in some cases cause tangible harm (being labelled a racist, a homophobe or a transphobe may inflict social stigma or result in job loss etc.) it is not the counter speech specifically that is viewed as the cause of the 'punishment', but the original transgression that invoked the counter speech. As such, social media platforms have no particular policing mechanisms in place to control its use, as they would hate speech (unless, of course, hate speech is used as counter-speech).

In their investigation of counter speech, Richards and Calvert (2000: 553) discuss the "doctrine of counter speech", as first suggested during an American legal case by Justice Louis Brandeis. In his statement he suggests that "If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the process of education, the remedy to be applied is more speech, not enforced silence". In this quote and in Richards and Calvert's (2000) research more broadly, ideas of "falsehood and fallacies", "bad speech" and "hate speech" are discussed interchangeably, with the idea of additional countering speech always preferable, where possible, to outright censorship. While the motivation to avoid censorship can be both ideological (free speech is sacred, no speech should be illegal) or logistical (not all "bad" speech can be effectively policed), the preference and utility of counter speech as a means to oppose hate speech occurs frequently among much of the research that discusses it.

In his work on counter speech, Williams (2019: 40) defines the practice as "any direct or general response to hateful or harmful speech which seeks to undermine it", before expounding its advantages over official legal sanctions. In dubbing counter speakers the "first responders" of a hate speech incident, Williams (2021:228) suggests that the speed of its deployment, its situational adaptability and its ease of use by any untrained internet user make counter speech a useful, effective and convenient alternative to censorious intervention by platforms or punitive response from police. He also suggests that the effectiveness of counter speech is at its utmost when performed in groups rather than individual intervention. While Richards and Calvert (2000) acclaim counter speech for its

avoidance of infringing on the American first amendment right to free speech, Williams' (2019) discussion focuses far more on the logistic advantages of counter speech, as a tool universally available to social media users, to wield as and when they feel necessary. Wright et al. (2017: 57) reiterate many of these advantages when they define counter speech as a "direct response to hateful or harmful speech [which] can be practiced by almost anyone, requiring neither law nor institutions". While Wright et al. suggest that "institutions" are not a necessity for the deployment of counter speech, it should be noted that institutions do engage in counter speech, whether official, corporate or charitable in nature. The recurrent theme of counter speech as a strategy free from official punitive powers also highlights notions of community policing, social responses to crime and deviance, and rehabilitative rather than retributive interventions, as promoted by Braithwaite (1989) and others.

Ernst et al. (2017: 7) define counter speech as any speech which "aim[s] at challenging these transmitted ideas of hatred, prejudice or even extremism" and posits the idea of a "counter narrative spectrum" in which the broad strategies for creating counter speech can range from "alternative narratives transmitting values of tolerance or freedom, up to counter narratives which de-construct and challenge extremist ideologies". This explanation develops upon previous definitions, describing not only the abstract communicative aims of counter speech, but providing broad-stroke examples of the kinds of narrative responses these kinds of speech may take; challenging the logic or validity of hateful claims, or evangelising positive values as an alternative to hateful ones. Schieb and Preuss (2016: 5) work in the same vein, providing generic strategy examples in their counter speech definition, suggesting that "reasonable and accurate arguments, facts and figures, employed in direct response to hate posts, are seen as a helpful treatment to restrict the impact of hate speech". The emphasis here suggests that logical and information-based responses are the most useful strategies for combatting hate speech and stemming its propagation.

While the specifics of the techniques used to deploy counter speech may vary, each of these definitions agree that the key communicative function of counter speech is to work to undermine hate speech, invalidate the stated positions of that speech, and reduce its impact in whatever way possible. Counter speech is not merely a response to illocutionary violence with opposing violence, but it is a rhetorical process which reduces the validity of the

hate speech and/or speaker and impedes the propagation of future hate speech. Williams (2019: 40) suggests that the undermining or invalidating process in counter speech can lead to two desirable outcomes; the first being the reformation and reduction of future hate speech just mentioned, and the second being an impact on the wider audience by "communicating norms that make hate speech social unacceptable or by 'inoculating' the audience against the speech". This secondary outcome is important not only because it reinforces the social norm to other members that hate speech is unacceptable and abhorrent behaviour, but also illuminates the protection victims may feel, seeing speech that may target and dehumanize them being rejected and retaliated against by other community members.

Richards and Calvert (2000: 559) summarise what they suggest are the three key reasons for counter speech's effectiveness by recounting an instance in which Ku Klux Klan (KKK) signs were erected and subsequently responded to with counter speech. Firstly, they suggest that counter speech was an effective response to the KKK because it upheld the importance of "freedom of speech" . They note the unavoidable irony of being "tolerant of intolerant expression" by allowing hateful and racist signs to remain, but again refer back to the "doctrine of counter speech", which asserts that whenever possible bad speech should be met with more speech, not censorship. Beyond the discussion of the right or wrong of hate speech specifically, in protecting freedom of speech in this way this stance allows for the protection of minority viewpoints that may experience censorship not because they are inherently aberrant, but because they are not held by the ruling majority. Secondly, Richards and Calvert believe that the audience beyond the interaction, outside of the hate speaker and the countering respondent, can experience a form of "self-realization and self-fulfilment" in seeing virtuous ideals that they share being represented and achieved in public. The "therapeutic" and "healing" aspect of seeing hateful behaviour rebuffed in public, Richards and Calvert suggest, is an important part of the process of counter speech. Counter speech works not only to challenge, punish and reform the offender, but also to restore and reconstitute those who are victimised and those who take offense to those aberrant views. Lastly, the public nature of counter speech helps to broadcast the message and norms delivered in the process, but also the mechanism of the process to a wide audience. In viewing counter speech, others will not only observe and internalise the notion that hateful speech is wrong, but also come to understand the process by which it can be opposed, were they ever

to come across it again. These secondary, echoing effects of counter speech on those *around* the interaction are an important part of the communicative aims of counter speech, even if they may not be the primary driver for engagement.

With the working definitions of both hate and counter speech laid out, this chapter now shifts its focus to the criminological aspects of online study, discussing the pertinent crime control ideas as informed by the scope of networked society and the unique opportunities and challenges afforded by its defining features.

## The New Criminologies of the Digital

Garland's seminal work on The Culture of Control (2001) describes and analyses the important shifts in focus that occurred in crime control following the post war period to the modern day as well as the changing technological, socio-economic and socio-political landscape that influenced it. In conjunction with Castells' Network Society (1996) and its forward-thinking view of technological progress through Computer Mediated Communication (CMC) as well as Scheff (1987) and Braithewaite's (1989) work on shame and its re-integrative properties in crime reduction and prevention, this section aims to explore and synthesise these ideas into a cohesive picture of the current state of digital crime, criminology, as well as the most prevalent and predominant theories on its responses.

One key concept that appears particularly applicable in the modern understanding of the crime problem is what Garland (2001: 128) termed "the new criminologies of the everyday" and their basis in the classicist theories of rational choice (Clarke and Cornish, 1987) and routine activities (Cohen and Felson, 1979). Garland posits that technological advancement and the social developments of late modernity have provided a cultural landscape in which the classic situational controls which shaped interaction and curtailed social deviance have reduced, whilst providing more opportunities and targets for crime. Considering the rationalist, classicist thinking that underscores this idea, with greater opportunity and less risk comes a

heightened likelihood that the rational actor will act in their best interest, which is personal gain.

These classicist criminologies are often also associated with situation preventative measures which pro-actively respond to possible crime. By installing preventative measures, risks are increased and the calculation to act criminally is stacked towards desistence. Mooney (2000: 17) suggests that these preventative measures typically take the form of "better locks and bolts on houses, improving environmental and architectural design, creating defensible spaces, Neighbourhood Watch, and so on". These can be seen mirrored in digital crime prevention, where algorithmic content restriction invokes the suggested architectural elements, and socially enforced norms and counter speech provide the work of a digital neighbourhood watch.

This has a clear parallel to the development of the digital online world which so ubiquitously pervades our current cultural landscape. The "digital" itself is now an everyday arena for social interaction, as well as work, recreation, shopping, consumption and almost any other conceivable activity. While the opportunities afforded to the digital citizen are expanding and permeating social life, so are the opportunities for 'digital deviants' to exploit new avenues for crime.

While Garland focuses on the physical changes that occurred allowing greater opportunity for crime, Adam (2003) and Bauman (2007) approach the technological advancement through a temporal lens. For them, one of the predominant factors to note in the technological change concerns the speed at which interaction can occur. According to Adam, the sequential delay between thought, action and reaction have become collapsed to the point of near instance using computer mediated communication and global digital transaction. This speeding up or 'collapsing' of time is viewed as an ameliorating good as the modern world has commodified time, and the quicker an interaction can be achieved the more time interactants have to perform other actions or consumptions. This speaks to the old idiom that "time is money", but in the modern digital age not only is time commodified but almost every facet of digital life has monetary value to be exploited. Digital lives, almost in their entirety can be

commodified. Not only are the digital venues and arenas of interaction private businesses run for profit, but the interactions themselves are being sold to third party companies and investors to better target consumers. The "time is money" commodification presumes thankfulness on the part of the user for expediating their experience, therefore saving money, shortening time on thought and reducing the need for patience and waiting, but this also speeds up the process of creating more interactional data and metadata which can be sold. Not only has the time in which you operate and the arena in which you digitally stand been commodified, but so have the digital actions of the user. This illustrates the extent to which almost all aspects of digital life can be exploited for personal gain, from your data, to your safety, to more classically understood forms of crime transposed online, such as digital violence and robbery.

As well as commodified time providing further opportunities for exploitation, it also causes a dematerialising and a decontextualising of the information that is passed, relinquishing control of that information from user to user. This disassociation leads to a lack of solid network bond across those interactions, and falls into what Bauman termed a "wish fulfilment" culture, accompanying a setting in which the processes of consumption and the goal of consumption become instantaneous. When applying these ideas to deviance on social media, particularly hate speech and counter speech, obvious parallels appear between the lack of ownership or permanence of the interaction for the sender and the ease with which hate speech is pronounced without fear of repercussion. This also lends itself to the discussion on the speed of a response that counter speech necessitates, as well as the effect that such response will have for either interactant. It can be assumed that for any response to hate crime or digital deviance to have an impact it must occur with the Beccarian (1764) factors of Certainty, Severity and Celerity. As a result of the vastness of online deviance certainty and severity cannot always be assured, so the speed with which punishment or response is applied can be seen to have an important deterrent effect in an instantaneous forum. Evidently, an instantaneous response to all online deviance is impossible, particularly when that response is expected to be provided by an official authority, such as a policing structure. This leads to a particular need for expanded informal social controls as an augmentation to traditional responses.

Informal social controls, and the tools with which they are achieved, are embedded throughout the digital landscape. From design choices in site layouts to available moderation tools, there are a host of situational techniques that empower informal social control. Along with these tools and designs, for an effective informal control structure to grow amongst its user base, users must themselves feel a connection to the community and 'social capital' (Bourdieu, 1986) to invest time and effort into the maintenance of that community and its norms. Castells (1996) notes that during the initial uptake of the internet, while it was swift, it was not universal and those who found themselves disadvantaged due to class, race, region or economy were less likely to be early adopters. This is important to note because while the internet is established and much more universal in its reach now, during its advent the internet and the world wide web were open source and often shaped by its users. Technical knowledge and capability allowed early users to create and form communities, organise socially and record and archive information as they wanted. The "hashtag" is an example of how even non-skilled but early platform users could influence the way in which platforms and types of social communication were organised and the effect users had on shaping their own environments. When users can see the impact and influence they wield over their community, they are more closely tied to and invested in the online space. This may well provide greater motivation to intervene with forms of informal social control, and to also take heed of warning or intervention by members of their community, rather than those of a more classical, 'looming' authority.

To further consider control in online spaces, Garland's (2001:124) notion of "responsibilization" is useful. This concept suggests that rather than by deputising citizens through state run or authority led community policing, responsibilization seeks to empower citizens to perform social control independently. This concept dismisses the hierarchy of citizen and authority figure in favour of programmes which instil civic responsibility and accountability to citizens and institutions to reinforce social norms and report on deviance and low-level issues. Deputised responsibilization appears ubiquitous in online social media, as almost all platforms provide some tools by which the community can monitor and moderate the speech of its citizens. These include like/dislike buttons, reporting mechanisms, community moderators, and community standards and guidelines which educate users about

how they should behave and what behaviours they should not tolerate in others. Implementing this type of scheme in the terrestrial world necessitates motivating a community to 'responsibilized' action when there is already an established and expected crime control infrastructure in place which the public trust to provide this service. Peeters (2013:538) specifically addresses this idea in his discussion of Dutch political discourse around "producing public value" and "nudging" its citizenry.

Hinds and Grabosky (2010) look at the individual factors which are most likely to instil personal responsibility, those being the expectations of police attendance and the satisfaction with performance, as well as differences of gender, education and fear of crime. This issue does not present itself in the online social control debate because no specific policing institution was installed, there was only an assumption that the terrestrial police would handle any crimes that were deemed serious enough to warrant a "real world" response. Button and Whittaker (2021) discuss what they perceive as a "lag" in state response to catch up with the scope and volume of online criminality, and the tendency for online citizens to accept voluntary measures, private policing and vigilantism as a replacement. Due to this lack of established digital crime control structure throughout the duration of the internet's 'lifespan', a notable portion of internet users oppose the idea of intrusive regulations and authoritative structures in favour of the anonymity, privacy and liberty afforded them. Garland suggests that in the terrestrial realm "sovereign" style criminal justice governance is gradually being replaced by something closer to Foucault's "governmentality" (Burchell, Gordon and Miller, 1991). Garland describes governmentality as "a modality that involves the enlistment of others, the shaping of incentives, and the creation of new forms of co-operative action". As a result of the lack of established sovereign criminal justice system in online spaces, control and moderations have been achieved in part through collaboration with professional and non-state institutions and the incentive for its citizens to be 'responsibilized' to social action for continued use of the private platforms and online services they enjoy in their everyday lives. Counter speech particularly fits the mould of these incentivised systems of co-operative action.

In order to implement democratised social control among the online population, mass surveillance as a motivating and regulating influence would be undesirable due to the scope

of what would need to be monitored. Instead, Garland advocates a programme of cultural intervention and a core alteration of the beliefs and expectations of standards within the community held by its members. Garland (2001: 126) suggests that "…the most important processes in producing order and conformity are mainstream social processes, located within the institutions of civil society, not the uncertain threat of legal sanction". Here Garland advocates for the idea that deterrence from deviance is not brought about by the threat of punishment, bound up with all of its uncertainty and inconsistency, but by social norms applied and reinforced through social institutions and community influence, e.g. counter speech. For lasting change to be achieved it must come from within the society and the digital communities, defined as they are by their multiplicity and oxymoronic immediacy and distance, are more effectively controlled by the citizenry than they are a punitive authority.

## Digital Situational Crime Prevention

The preference towards informal and democratized social control online can be seen in the consistent application across social media platforms of user focused moderation tools. Users across most, if not all, social media platforms are given the responsibility to perform low level content moderation, through integrated functions like report buttons, "likes" and upvote/downvote buttons. While these are not immediate sanctions upon a deviant user and may require additional intervention from the platform, these act as community maintenance tools to help users enforce the standards and norms they see fit for their digital community. These tools not only provide real time normative feedback to users, but also in the case of platforms like Reddit, can collapse content out of view when a certain negative threshold is achieved. These low-level community intervention techniques harken back to Wilson and Kelling's (1982) broken windows theory, suggesting that when low level transgressions are attended to or prevented, more serious acts of deviance are less likely to occur. Fredheim et al. (2015: 1) touch on this in their discussion of anonymity in online comment spaces, where they note that when "trolls and spammers" are excluded from the community the overall quality of interaction increases.

Garland (2000, 2001) is once again relevant here, in his discussion of the modification of usual practice through the intervention of multiple agencies and non-state actors. In general, most of the work done policing the norms and culture of online spaces is provided by non-state actors, be it users or the private company run platforms themselves. The platforms provide an online space for community and communication, but are also the key stakeholder in policing, or providing the tools for policing, those digital spaces and publics. Reddit and other sites who are organised around particular sub-communities delegate their power and deputise users to enforce their chosen standards, often without much oversight from the platform themselves. The idea of multiple agency influence in crime and its control are also paralleled by Shearing and Wood's (2003) discussion of nodal governance. They suggest that while a plurality of institutions are involved in the development of new forms of governance and policing, they are connected and influenced by one another in important ways. As an example, they discuss the private sector innovation of proactive preventative crime control techniques which have filtered back into traditional policing. Private online platforms so too have used their ample resources to research and develop crime control methods that have influenced or been adopted by state policing institutions.

Chandrasekharan et al. (2017) investigated one of the few instances where the platform did intervene and wield their policing power, examining Reddit's decision to delete two communities created around fat shaming and anti-black racism in 2015. The study sought to identify if the ultimate sanctioning of these communities effectively reduced hate speech or if that reduction was mitigated through migration or transformation, invoking the proselytization and proceeding criticism of hotspot policing by Sherman (2009), Carr (2010), Hope (2009) and Tilley (2009). Although Chandrasekharan et al. (2017: 31:15) could not track the possibility of migration outside of Reddit, they found that within the platform a "large, significant percentage of treatment users from the band communities left Reddit" and did not infect the discourse of other communities.

With this discussion of the developing landscape of criminology in mind, this chapter will now turn to an overview of the digital arena that these new criminologies seek to affect. Social media, as will be discussed, has had an exponential impact on society, not only in its prevalence and ubiquity, but in its influence on the ways in which users interact. In this next

section, the architectural, networked and spatial-temporal features which typify this new form of online communication will be unpacked and examined to provide an understanding of why Garland's theories of control and responsibilization fit so usefully within its parameters.

## THE "SOCIAL" OF SOCIAL MEDIA

Social media websites in all their permutations, such as Facebook, Twitter, Gab and YouTube, are archetypal examples of what has come to be discussed as 'Web 2.0', a version of the internet defined by "communication, user-generated content, data sharing and community building" as opposed to one primarily designed around the delivery of information (Fuchs, 2011:288). In O'Reilly's (2005) definition, he stresses the architectural focus of this new era of online services encouraging participation and interconnectedness across both devices and platforms. Social media and its networking features have been integrated thoroughly into other online services (banking, bills, shopping, etc.) and into offline spaces (hashtags on news programs, QR codes on bus stops, etc.) and as such more interaction is funnelled through the filter of digital social media.

However, although online interaction is becoming more ubiquitous, and despite the technological developments in the multimodality of those interactions, there are nevertheless important, definitive distinctions between "computer-mediated-communication" (Herring, 1996), or CMC, and its offline counterpart. These essential differences alter not only the way we achieve the aims of our communication but the impact that those forms of communication can have. The digital interactional architecture of different platforms constrains communication in different ways, providing users with a different set of interactional tools and rules to work with, all of which are implemented through the distanced, anonymous and temporally muddied filter of CMC.

### Distance, Anonymity, A-synchronicity and Digital Etiquette

Anonymity online, and its effects in increasing aggression and deviance in communication, is one of the central structural tenants of social media most often raised in popular discussion (Rainie, Anderson and Albright, 2017; Chiles, 2019; Phillips and Bartlett, 2018). 'Cyber bullying', 'keyboard warriors', 'flaming', 'trolling' and many other new additions to the common lexicon all stem from the notion that the anonymity of digitally distanced communication affects interpersonal interaction online. While anonymity is not always total, almost all platforms provide anonymising, distancing features which are perceived as negatively affecting interaction. Kiesler et al. (1984) have compiled a list of features inherent in CMC which they suggest have a social-psychological impact on the ways people understand online communication and impact the way they go about their interactions. While these foundational ideas have provided a springboard for more contemporary research in the developing and changing landscape of CMC, they too are worth considering here to provide a useful base of understanding to this research.

The first foundational aspect of CMC which Kiesler et al. (1984: 1125) discuss in contrast to face to face communication is the "absence of regulating feedback". They posit that "head nods, smiles, eye contact, distance, tone of voice…" are among the wealth of non-verbal, co-ordinating feedback, which in offline communication work to maintain the flow of conversation and regulate the power dynamics within interaction. These prosodies work to maintain effective communication and are very difficult to properly portray through text only communication, even with the additional multimodality afforded by image, emoji and other communicative aids. This lack of feedback contributes to the a-synchronicity observed within online communication, where the turn taking structure often presumed to be naturalistic in face-to-face, or even voice only communication dissolves when not propped up by recurring non-verbal interactive markers. These non-verbal cues all work to demonstrate consideration, politeness and proper social etiquette in a face-to-face exchange, and are often totally absent in CMC. Platform designers have attempted to address this deprivation of non-verbal feedback through the development of multimodal communication and "typing awareness indicators" (Stenovec, 2017). However, in general users find themselves without the means to properly portray tone or accurately understand the turn taking structure. The attempt to bridge this particular gap in interaction suggests that its absence has a noticeable negative effect on the communication people are able to achieve online.

Kiesler et al. interestingly frame this discussion of regulating feedback in terms of the deindividuation and a loss of self. They suggest that computers tend to be "absorbing and conducive to quick response", which weakens normative self-regulation, leaving greater opportunity to engage in interactive behaviours outside of the norm. Additionally, they posit that deindividuation of this kind is often coupled with a lack of hierarchy or status signifiers. Platforms such as Twitter do contain a verification system, where check marks are assigned to users who are of public interest in the realms of "music, acting, fashion, government, politics, religion, journalism, media, sports, business and other key interest areas" (Twitter, 2017), however the impact of these check marks may pale in comparison to other offline indicators. Kiesler et al. note that students were more confident to ask questions of their professors through email than face-to-face. This would suggest that in spite of verification check marks people are prone to 'trolling' or antagonistic speech with celebrities or people of high status more frequently than they would offline.

Baym (2010) offers a contrasting, but complimentary, interpretation of the influence of a lack of prosodic cues on interaction. She suggests that while the lack of cues may on the surface appear to ease the user's ability to fabricate their identity, heightened anonymity simultaneously reduces the likelihood of consequences for honesty that may transgress expected social norms. And with this lack of consequences comes a greater freedom to engage honestly in self-expression through online interaction, which may in itself offer some explanation as to why antagonistic speech seems so much more readily given when compared to its visceral, feedback laden offline counterpart.

In a study of group consensus finding in face-to-face, computer mediated, and anonymous computer mediated communication, Kiesler et al. found that computer mediated groups were more uninhibited than face-to-face groups and more prone to swearing, insults and hostile comments. They were also found to be more unwilling to back down from their opinions and as a result took longer to complete tasks. Kiesler et al. described the level of online etiquette as "immature", suggesting that because of the distinct subculture of early online adopters, conventional social boundaries were blurred, familiar or harsh language was found in professional settings and privacy concerns often flaunted. While left over artefacts

from the advent of social media still prevail, shared norms and novel practices for structuring online interaction have developed. Most prominently the “hashtag”, which was originally developed as a user-created archiving and searching tool, has been adopted across many online platforms and is proliferated among classical media and popular spoken language. While the development of this technology appears useful and convenient for users, it is not without the possibility to draw people sharing hateful ideologies together, or to draw those hateful ideologies to groups who may be unwitting targets in online discourse. More-over, this kind of language search function can act as a methodological tool for discovering and addressing deviant speech.

Khan (2017), in his comparison of engagement on different platforms, notes that lower levels of anonymity produce ‘higher quality’ comments and interactions on YouTube, and Baron (2008: 101) posits that anonymity online allows “Those either so lowly they ought not to presume to rise, or so high that they should not have sunk, to involve themselves in public debate”. This would suggest then that the anonymity and distance provided by CMC, whether for good or ill, encourages more people to interact in ways and situations that they would not in offline areas. The confluence of these technologies can be seen to result in interaction that is more unregulated and unmediated, but that is also encouraged by platform design and anonymity. With the removal of social norms and restraints, users may feel emboldened not only to engage in unacceptable hate speech but may also feel emboldened in combatting it.

## The Theories Driving This Analysis

### Digital Vigilantism, Identity and Shame

As noted, the incredible scope of online crime means that an effective official response is unlikely, and so that responsibility must fall on the plurality of private institutions and community members. Those community members, emboldened by the tools provided by platforms, are afforded discretion in how much they choose to intervene. Some watch from the side-lines, while some go as far as to initiate accounts and pages solely focused on

vigilante activity, exposing and shaming those who transgress the approved norms of online spaces. “Sentinel” sites (Webb et al. 2016) such as @YesYoureRacist, @RacismWatchdog, @Homophobes, have proliferated social media with the express purpose of providing vigilante justice through exposure and condemnation of deviant online behaviour. Sentinel sites are accounts, normally centred on one specific social justice cause, who re-publish social media posts which they find to be hateful. This broadcast shaming works both as a direct punishment of deviance and as a more generalised group deterrent to those who view the wrath of online shaming. Again, this deterrence theory harks back to the classicist criminology of rational choice and routine activities, as well as drawing on notions of the synoptic (Mathiesen, 1997) and the panoptic (Foucault, 1995), enforcing and internalizing behavioural norms through surveillance and expected threat of retribution or punishment.

A great deal of the deterrence provided by socialised community controls, through surveillance, moderation and counter speech, is empowered by the idea of shame and its impact on personal identity. The power and use of online identity will be examined through the data in its dedicated analytic chapters (chapters 5 &8) within this thesis, however first it is important to discuss the theory of identity that will frame its understanding. This research takes as a core tenant the symbolic interactionist theory of identity as pioneered by Mead (1934) and Blumer (1969) and developed by Goffman (1959, 1961a, 1961b, 1963, 1967). A defining principle of symbolic interactionism holds that identity can be understood to stand distinct from any essentialist notions, and instead is formulated, maintained and performed through interaction. Identity is not something that one has, but something that emerges and is curated in situ, through the text and speech of interaction. Nowhere is the fluidity and transformative nature of identity more apparent than in online social interaction. Online social interaction strips back a great deal of the physical and contextual factors that can influence how identity is understood and characterised, and instead the focus of identity performance is boiled down to the pure text of interaction. While there are additional multi-modal identity markers that can be accessed through online interaction (images, emoji, symbols, biography details), this examination takes as its focus only the identity work that is done *in text* by users. Goffman’s dramaturgical analogy and particularly his conception of “face” (Goffman, 1967) provides a useful base for interpreting much of the rhetorical work

that is done in online interaction, as well as a frame for viewing the importance of shame and shaming in maintaining interactional life.

For Goffman, the "face" is something to be preserved and protected from damage, something that can occur when deviance from the "line" brings stigma and shame upon the person. West and Trester (2013) apply the notion of face and facework to online social media directly, exploring how Brown and Levinson's (1978) conceptions of positive and negative face are attended to through interactional turns such as tagging photos (p.139) and status update commenting (p.142) on Facebook. In Asylums (1961a:20), Goffman discusses the use of "identity kits", tools and supplies individuals require to properly perform their identity. While online interaction is markedly different to life in a total institution, there are both demographic (profiles information) and interactional identity kits that are provided and actively used to maintain face and identity, and further to mitigate shame and stigma when possible. Shame and embarrassment are key drivers for proper performance of face and its repair in case of incident or "spoiled identity" (Goffman, 1963). To emphasise the importance of social pressure in the avoidance of deviance from the norm, Goffman (1961b:72) notes that "…to be awkward or unkempt, to talk or more wrongly, is to be a dangerous giant, a destroyer of worlds".

Goffman suggests that the shame and discomfort felt by a breaking of face is not restricted only to the individual whose face has been compromised, but all those in the interaction *and* the audience to the interaction. This may in turn prompt an attempt at face repair from either the damaged individual or another individual observing the deviation, providing a useful explanation for why those not specifically targeted in a hateful attack may feel compelled to intervene or counter. Shame and face damage are not only a consequence of deviation from interactional norms but are a driver to repair. The reception of this attempt at repair, or the "corrective process" (1967: 305), Goffman suggests, can be taken in some cases with gratitude, or in others refuted and met with aggression, as the shame and embarrassment is exacerbated by the outside attempt at repair. This idea too is taken up in criminological application by Scheff's (1987) idea of "shame rage spirals", explaining that while shame is a powerful behavioural motivator, its reception is vulnerable to myriad responses.

Scheff (1987) and Lewis (1971) both focus their sights on the psychological impacts of shame, discussing the internalised mechanisms that can lead from shaming to negative internal feedbacks such as shame-rage "feelings traps" and "humiliated fury" (1971:127), as well as "narcissistic rage" (Kohut, 1977). These internalised and escalating loops can lead shaming practices to enhanced anger and deviant behaviour because shaming is deployed without any positive, re-integrative avenue through which to deal with their shame. Braithwaite (1989) specifically frames this in a criminological fashion, discerning the difference between stigmatizing or non-re-integrative shaming, which targets the offender or deviant and shames them as a person, and re-integrative shaming, which more often focuses on condemning the deviant action. Braithwaite suggests that stigmatizing shaming of the person removes agency in the individual and reduces them to the deviant label they have been assigned, internalising a lack of opportunity to move beyond their misdeeds and re-join society as a productive member. Re-integrative shaming, however, seeks to condemn actions and provide the offender with the chance to rebuild attachments with the community in the hopes of preventing further recidivism.

Connecting Goffman's sociological and symbolic interactionism ideas of face with Braithwaite and Scheff's criminological applications is not done here to restate one's influence on the other, but to illustrate the idea that the criminological conception of these ideas is and can be used in a weaponised form. While Goffman identified shame or damage to face as mostly a consequence of interaction or internal motivator, the criminological application often transforms those ideas into something that can be deployed upon a person to achieve the social goal of effecting behaviour.

Sykes and Matza (1957) provide an alternate but complimentary view on shame and its internalised motivation to offenders. They base their analysis around the study of delinquent offenders, but some of their key points seem to ring true with regards to online hate speakers as well. In their study on "Techniques of Neutralization", Sykes and Matza (1957: 665) suggest that rather than holding a completely separate set of moral values as a way to justify their behaviour, delinquents oxymoronically are able to hold a particular set of deviant norms while "recognizing the validity of the dominant normative system in many

instances". Delinquents understand that their normative structure exists within the schemes of the larger world and uses that interplay to discern right from wrong within their own flexible morality. Some victims are 'appropriate' and acceptable targets, while some are not. Some deviant behaviours are justifiable, while others are not. Sykes and Matza (1957:666) suggest that these rationalizations are "viewed as following deviant behaviour and as protecting the individual from self-blame and the blame of others after the act" but also propose that justifications can "precede deviant behaviour and make deviant behaviour possible". That is to say that deviants are aware of the conflict between the morality needed to complete their actions and the norms of the general society, and the shame generated in that conflict leads them to produce rationalizations in certain instances to offset their internalized guilt and carry out their deviance.

Sykes and Matza propose 5 key techniques by which a delinquent may neutralize their internalised shame and justify their aberrant behaviour: denial of responsibility, denial of injury, denial of the victim, condemnation of the condemners and an appeal to higher loyalties.

Denial of responsibility is characterised by the proposal that the agency of the person to commit the act is removed, and the responsibility for the deviant behaviour is the result of accident, bad upbringing, poor influence or suspect neighbourhood etc. In the context of online hate speech, a discussion that engages in racial stereotypes may not be intended as harmful, but someone may take it as such. If a person accidentally employs racist stereotypes but did not mean to cause damage by them, they have no responsibility for harm and it is therefore justified.

Denial of injury requires the delinquent to make the distinction between acts that are immoral because of their harm and acts that are illegal despite a lack of injury. If the delinquent believes no harm has occurred and it is only arbitrary rules that have been broken, they can neutralize any shame that may impinge on their action. If they believe a slur, or linguistic violence in general, provides no actual harm to a person, then they will have no moral issue braking capricious rules.

Denial of the victim is based around the circumstances of the event nullifying the recipient of damage as a proper victim and remakes the violence as a retaliation or punishment. Sykes and Matza (1957: 668) themselves reference attacks on homosexuals and minority groups who may have gotten "out of place" as exemplifying this technique. In terms of the anti-immigrant hatred discussed in parts of this analysis, if immigrants are believed to be entering the country under false pretences, to attack traditions or 'take' undeservingly from the welfare state, then their demonization and dehumanisation does not make them a victim, it makes them recipients of retaliatory justice.

Condemnation of the condemners appears particularly apt for the discussion of online hate speech and counter speech, as it involves a deviant bypassing internal reflection of their own behaviour and focuses their attention on undermining the motives of the disapprover, suggesting them to be "hypocrites, deviants in disguise, or impelled by personal spite" (p.668). The recurrent weaponization of "white guilt", "SJW's" and "Performative wokeness" as means of decrying the work of counter speakers in popular culture speaks to the power and pervasiveness of this deflecting tactic as a means of neutralizing shame.

Finally, appeal to higher loyalties purports that known deviance can be performed despite social and legal pressures because of the allegiance with a group seen to be more important in the eyes of the deviant. A fidelity to race, nationality, religion, political alliance, family, friendship or gang may all be strong enough for someone to justify deviation from the social norms at large. Sykes and Matza suggest that this or any other technique, displays how deviant actors suffer from internalised guilt and shame because of their awareness of the norms they are transgressing, but are able to justify through a more pressing or relevant norms.

In an online setting, these conceptualisations of identity and shaming seem particularly applicable. Identity, beyond being a core deciding factor in targeting victims of hate speech, is something that is constantly and consistently discussed and weaponised in online debate. The digital dramaturgical divide between the offline and online means the that the performance of identity is a continuous and evolving thing, moving decidedly beyond "bios" and "profiles" and into the everyday text interaction to provide users with personhood,

authority and authenticity. Each of which are at constant tension with the possibility of being shamed or shown to be performing badly. Shame is not only an internalised influence, prescribing how social identity is performed, but is a key feature of counter speech, one of the primary social tools for responding to online hate. Importantly however, the way in which shame is deployed through counter speech matters greatly to its rehabilitative or re-integrative outcomes. Counter speech is often touted as a universal tool for everyday users to positively influence the communities they live and interact in, but the rhetoric used in that counter speech can have different and contrasting impacts to those receiving it.

This chapter will now turn to a discussion of Swales' Genre Analysis (1990) to establish a definition of "Genre" as a term for the categorisation of discourse and explain how hate and counter speech fit into this definitional framework, before problematising that framework through a constructionist, intertextual lens.

## The "Genre" of Hate and Counter Speech

This final part of the literature review will discuss Swales' notion of 'genre' as a method of linguistic categorisation. Swales' genre analysis will drive the investigation into both hate and counter speech as an analytic frame work in itself, as well as providing a base understanding for how these discourses are understood and engaged with throughout this thesis.

Swales' conceptualisation of genre is achieved through a few disciplinary lenses, touching on literary studies, linguistics and rhetoric. However, despite this range of applications, he outlines five criteria by which he believes a genre can be identified. Each of these criteria will be outlined and discussed with reference to how online hate and counter speech meet the demands of those criteria. In designing his conceptualisation, Swales draws frequently on Miller's (1984: 155) work on the genres of rhetoric, who suggests that such homely discourses as the "progress report, the ransom note, the lecture…the eulogy, the apologia, the inaugural, the public proceeding, and the sermon" do not trivialise or demean genre as an area of study, but that embracing these categories of discourse instead takes seriously the rhetoric with which people find themselves in contact every day. It is this

chapter's function to identify online hate and counter speech as generic rhetorical categories (as "genres" themselves) and treat them with the seriousness that their content and impact demands.

Swales states that first, genre is a *"class of communicative events"* (Swales 1990). This initial criterion is a broad, wide-reaching condition, understanding that a genre must communicate something, whether by text, speech or through non-verbal means. Swales also notes here that a genre of communicative event can range in the frequency of its occurrence, from common everyday communications, to rare or occasional, and that the event is comprised of "not only the discourse itself and its participants, but also the role of that discourse and the environment of its production and reception, including its historical and cultural associations" (p.46). This foregrounds Swales' and others later assertion, that while genre as a classifier suggests a closed and rigid system by which to categorise discourse, it is in fact created in communication with a host of influencing factors beyond the form and structure of its bare text. Hate and counter speech fall within this framework, each comprising a class of communicative event, communicating either a specific kind of hateful and incendiary language, or language that combats and seeks to influence behaviour, both of which are bound up with the socio-temporal associations of their contextual setting.

Swales' second stipulation is "*The principal criterial feature that turns a collection of communicative events into a genre is some shared set of communicative purposes."* (p.46). Here Swales clearly distinguishes the focus of genre as the purposive intention of that genre as opposed to its form and structure, noting that were this focus inverted, "facile" classifications based around "stylistic features" could be asserted as genre where it would be inappropriate. A genre is defined by its intention to achieve a communicative goal, and Swales goes further to state that a single genre may indeed contain a whole set of communicative purposes, all working concurrently. To illustrate the necessity of a focus on the purpose of communication over the structure, Swales invokes parody, noting how parody often deploys the surface level recreation of form and structure to derive absurdity and derision when those stylistic features are applied to contrastingly inappropriate subjects. Again, hate and counter speech meet this framework by their identifiable communicative purposes. They each have

aspects of form and structure that can be and are present, but what more readily defines them as their specific genres are their intended purpose as forms of speech action.

The third criterion is "*Exemplars or instances of genres vary in their prototypicality.*" (p.49). This criterion expands the range of what is expected to be typical within the assumed tropes or genre markers of any particular genre. Swales asserts that there are two approaches to identifying an utterance as part of a genre; "definitional" and "family resemblance". While identifying utterances based on their adherence to a definitional set of characteristics is a better-established model, Swales notes the inevitable issues of insufficiency with creating a definitive list of characteristics that would define in and out all those instances that should be included under a given genre. Swales exemplifies this issue, noting that it would be almost impossible to create a list of characteristics that could accurately identify everything as a joke, while inversely any accurate definition of a bird is likely to lose its precision when presented with a roasted chicken. In contrast to this, Swales suggests the "Family Resemblance" model, in which overlapping similarities and equivalences adequately signal the resemblance of different instances, identifying them as fitting under that genres label. It is a broader and more interpretive stance, and although not free from criticism (a familial resemblance between A and B, and B and C may imply a family resemblance between A and C when none is present), is one that is more receptive to the notion of providing generic prototypes for talk and text.

This second, familial model, suits the varied prototypicality of online hate and counter speech. While the form, structure and specific language may vary between hate and counter speech focused on different groups (attacking or defending marginalised races, religions, sexual orientations, genders, etc), there are clear resemblances between these differently focused communicative actions that house them under the banners of either hate or counter speech. Racist hate speech is identifiable as similar to transphobic hate speech, although the slurs and justifications used in each instance may differ, and speech defending immigrants is identifiable as similar to speech advocating for LGBTQIA+ rights, although its form and structure may differ.

The fourth of Swales' criteria is "*The rationale behind a genre establishes constraints on allowable contributions in terms of their content, positioning and form."* (p.52). Swales notes that the rationale provided by the communicative purpose of a genre establishes the constraints and conventions by which it orders itself, but suggests too that those conventions are malleable, evolving and subject to change even though they continue to exert influence over content. Swales illustrates this influence of rationale on convention by invoking the differences between "good news" and "bad news" in administrative correspondence. While both may adopt a shared space in terms of "textual environment and register" (p.53) their individual rationale, as influenced by their communicative purpose, renders them to different genre classifications. So too then it can be argued that hate and counter speech, although sitting in this case under the same "supra-generic assembly" of online discourses, are distinct from one another by the rationale of their purpose; one is to other, belittle and attack, one is to defend, protect and police. This is then in turn established by the constraints on content, form and expectation of interaction. Good news, Swales notes is categorised by its positive tone and the assumption of a welcome and affirmative response, whereas bad news is often categorised by a signalling that communication has ended. Similarly, hate speech is often portrayed in many fashions, as proclamations expectant of silent acceptance, passive confirmation, or incredulous retort, where counter speech seemingly always assumes active engagement and often an argumentative response.

Swales' fifth and final criterion is "*A discourse community's nomenclature for genres is an important source of insight"* (p.54). Swales here asserts that specific genre terms created and deployed by those routinely involved within a genre can give insight into the expertise held by those discourse communities. That is to say, that where recurring or common words and phrases used within a genre can point to the competency of that community member as a proponent or operator of that genre. Swales once again notes that these nomenclatures are generative and subject to change and evolution over time. While it may seem distasteful to describe them as such, slurs and well-known dog whistle phrases can be understood as expert terminology wielded by those most intimately engaged with the genre of hate speech. So too are the agreed upon terms to identify oneself and one's combatants in these arenas, for example describing oneself as a proponent of "free speech" against an opposing "SJW" (Social Justice Warrior). Similarly, recurrent phrases (triggering, safe spaces) and identity markers

(Racist, Islamophobe, TERF) wielded by counter speakers are understood to be community nomenclature, and therefore are a source of insight into that genre.

As Swales has repeated throughout his definition, those exemplars that express his criteria in each case are subject to historical and social influence and change, and as such are dynamic and fluid. The content and form, emblematic of a particular genre, may change over time, but the genre can still be identified. Genre is not a fixed and abstract categorisation, but something that is created by members and achieved.

However, while it is members whose actions achieve these discourse categories, Briggs and Bauman (1992: 144) note that "genre as a classificatory concept does not necessarily imply self-conscious attention to classification". Here one can understand that although an identified genre may appear to cast obvious constraint and typicality upon a discourse, members working within that genre may not be conscious of their adherence to them. They are not engaging in discourse that fits a genre because they are referencing it against a generic ideal, but because they are part of a discourse community whose archetypes hold influence, whether conscious or unconscious, over their communicative action. The designation of a genre is not a universal constant, but something that is co-constructed by members engaging within its boundaries, and those designations, although congregating around the same generic principles, may be coloured differently. What one analysis might term political messaging; another may ascribe to be propaganda. This analysis understands that those engaged in the analysed texts may not subscribe to the idea that they are performing racism, hate speech or even counter speech, and are in fact performing a set of communicative events borne from an entirely different ideology. This is a separate discussion on the perspective differences in categorisation between community members and between those members and those viewing from outside. However, for the analysis being conducted here, this study interprets these texts as being part of the hate speech or counter speech genres designated.

With these genres now established as the subject of enquiry, this chapter turns to problematising Genre as a classification of discourse, and discusses Briggs and Bauman's constructionist approach to genre, and the issues of intertextual gap.

## Problematising Genre and the Intertextual Gap

Briggs and Bauman's (1992) work, as briefly alluded to earlier, focuses in on the problem of genre as a taken for granted form of language categorisation. Although generic classification assumes a rigid set of rules by which they must adhere to be accepted as emblematic of that genre, naturally occurring speech is often more dynamic, improvisational and fuzzy than would be considered to fit within these constraints. Briggs and Bauman (1992: 132) illuminate this flaw when they state that "all of us know intuitively that generic classifications never quite work: an empirical residue that does not fit any clearly defined category—or, even worse, that falls into too many—is always left over". Here Briggs and Bauman touch upon the idea that the inherent subjectivity of decoding and assigning genre as a classification results in an inevitable ill-fitting of text to genre, or an interpretable fit into many genres at once.

This is a particularly interesting assessment when coupled with Hymes (1972: 65) suggestion that "it is heuristically important to proceed as though all speech has formal characteristics of some sort as manifestation of genres; and it may well be true". Hymes (1974) later went on to reconfigure and clarify this statement, noting that while it may be tempting to assume all verbal material is definable within a generic categorisation, he believes that may differ between discourse communities and that what was worthy of investigation was what falls within and without generic classification and why. Combining these two ideas then, one could assume that genre classification can be expanded into a great deal of discourse and linguistic material in theory, but that in practice there is almost always a degree of uncertainty and fuzziness in this framing.

The problem that arises here then, is that genre continues to be an identifiable classification of discourse, despite the deviations in naturally occurring text and talk from the seemingly rigid structures and forms that define a genre. Briggs and Bauman draw upon the paralleling studies of Sherzer (1983) and Duranti (1984) of "*ikar"* and "*lauga"* speech genres, to expand the definition for generic classification, to suggest that generic classification "cannot be accomplished by the examination of texts alone, but resides rather in the interaction between the organization of the discourse and the organization of the event in

which it is employed" (Briggs and Bauman, 1992: 142). This suggests that genre cannot be identified based purely around text, devoid of context, but is established through the interplay between that text and its contextual event. The forms and structures that can define a genre may occur in a different context and be identified as a different genre, much in the same way that deviations from the assumed forms and structures placed within a context and an event may inform and confirm that genre. The salient implication here is that in a context (social, historical or technological) where restrictions and constraints upon the form and structure of hate and counter speech may be enforced, that context may amend those deviations and solidify the observed genre classification. A genre such as hate speech, which is arguably most readily identified by its use of slurs, can still be identified when those slurs are restricted by the context, because of the interaction between that context and the text or talk being used. Similarly, when the extended and in-depth engagement with hate speakers that may readily define counter speech is restricted by the tacit or enforced constraints of social media interaction, that discourse can still be generically classified as counter speech, because of the context in which it is found.

Briggs and Bauman (1992: 146) suggest that there has been a distinct shift in focus in generic classification from mere text to the "social and poetic dimensions" of performance. To refer to Swales' first criterion for genre definition, can one identify the performance of a communicative event, when the text itself may not fit perfectly with the assumed forms of that genre? A performance can impress upon an audience the authenticity of a virtuosic command of the structures and communicative purpose of that genre to identify it as such, without a strict focus on the text's alignment within the generic form. If a hate speaker can "other" a vulnerable minority group, degrade an individual or aggrandize a privileged group at the expense of another without necessarily deploying the presumed text associated with those actions, the performance can still identify the hate speech genre. Correspondingly, if a performance combats hate speech effectively through some means other than the questioning or request for information often suggested in the definitions and instructions of counter speech (which will be discussed in greater detail below), that performance can be generically classified as counter speech, through this more encompassing orientation of genre analysis.

Drawing on a more critical approach to discourse analysis, Briggs and Bauman (1992: 147) go further to suggest that beyond simply illustrating a person's virtuosity within a genre, an invocation of genre creates an "indexical connection" that links this utterance to geographical settings, historical contexts or "distinct groups as defined by gender, age, social class, occupation" outside of the current moment. Not only are these utterances bounded with ideology and social implication, they also invoke the power and authority to recontextualise a genre into a new setting. If a genre is identifiable, that means the discourse to which this genre refers has existed at some point previous to this, and that historical and social context is what is being drawn upon to understand the genre deployed in this discursive setting. As Briggs and Bauman (1992: 148) state, "even when the content of the discourse lacks a clear textual precedent" the generic intertextuality between the historical and the contemporary draws power and identity into the utterance. Generic framing devices like "Once upon a time…", or in this case "I'm not racist but…" invoke a great deal of context, that influences the generic classification of what follows, even when the preceding text is new and different. Moreover, this invocation of generic intertextuality imbues the producer of the discourse with an assertion of authority to wield the genre as they do.

Authority of a genre can attach constraints to a discourse producer as much as it affords power and opportunity. The historical, social and ideological context brought up through generic identification of discourse implies and informs a great deal about the text being produced and the person producing it. When performing a religious ritual, the immense power and reverence that is intertextually connected to those forms and structures may imbed within its speaker the rights and powers of religious authority, or dependant on the context, may identify someone to be illegitimate and sacrilegious. In examples such as this, adherence to or deviation from strict generic forms have implications for the authority and power held by that genre of discourse. That level of adherence or deviation can be thought of as the "intertextual gap" (Briggs and Bauman: 1992: 149) between text and genre.

As noted earlier, it is a necessary assumption that generic discourse, as it occurs 'in the wild' will suffer from an intertextual gap. Rarely will deployed utterances within a speech genre map directly onto the forms and structures associated with that genre, leading to the initial criticism that the focus on the textual elements of genre analysis must be altered if its

use is ever to be considered useful. Yet while the intertextual gap is an inevitable and unconscious consequence of natural discourse creation, it can also be consciously deployed to mediate the power of those generic adherences. By increasing the intertextual gap, one allows for greater freedom of improvisation within the genre, and conversely, decreasing the intertextual gap more explicitly grounds the discourse within that genre. Through these manipulations, discourse producers working within a genre can call upon specific forms and structures to identify themselves and give authority to a genre of discourse. Similarly, through the avoidance of certain generic prototypicalities, one can distance oneself from the historical or social implications of a genre, while still operating within it.

Genres vary in their power to structure discourse. Some genres, like religious ritual, enforce a heterogeneous structure and order to their discourses, through recitation and procedural ceremony. Other, more conversational genres of speech provide for a greater breadth of improvisation and flexibility, while still containing textual, performative or purposive markers that can be identified as being within that genre (Briggs and Bauman, 1992: 156). That greater flexibility and disorderliness means that there is a greater opportunity for a naturally occurring intertextual gap, but also a greater range of freedom in which producers of discourse can deploy the manipulation of intertextual gaps as strategies for discursive power.

This conscious or unconscious creation of intertextual gap is what will be investigated in this thesis. Briggs and Bauman (1992: 163) note that "choices between intertextual strategies are ideologically motivated, and they are closely related to social, cultural, political economic, and historical factors", so whether intentional or not, intertextual gaps achieve something, and are influenced by their contexts. In the use of hate speech, a maximised intertextual gap between the text and the genre can distance the text far enough that it is not seen as explicitly hateful or racist, despite its achievement of the expected discursive purpose. Alternatively, strategies that seek to minimise the intertextual gap can act as dog whistles or identifying beacons to their assumed audience that they are indeed engaging in a hate speech communicative event, while still adhering to the restrictive language standards of the platform or the current social norms. Similarly, a linguistic turn which maximises the intertextual gap between the text and genre of counter speech may cloak the combative nature of counter speech, rendering it more acceptable and persuasive to their opponent. A

minimizing of that same gap may work to increase the visibility of the public shaming process inherent in social media counter speech.

## Jokes and Comedy

Humour is a linguistic technique often used in both off and online conversation which performs many rhetorical effects, not least of which is its ability to justify or smooth over the discussion of taboo subjects. Billig's foundational work (2001, 2005) into humour and hatred provides an insight into the linguistic work achieved by racist jokes, but has also been an interesting counterpoint to research exploring the disarming, community building aspect of seemingly offensive or poor taste jokes (Terrion and Ashforth, 2002).

Billig's work centres around the power of framing speech as "just a joke" as a justification for the deployment of racist stereotypes or the explicit use of hateful slurs. Under the guise of humour, hate speakers can pre-emptively justify their use of hateful language without the need to engage explicitly with it; whatever was deemed offensive was done to shock and entertain and not to harm. The clarion call of proclaiming hateful speech to be "just a joke" works not only to justify the speech but also to in some way delegitimise those who would be offended as not understanding or being too sensitive about what must have been an innocuous comment. This technique therefore increases the social acceptability of whatever is said, be it generally hateful to a group or specifically insulting to an individual.

In his work, Billig (2001) dismisses the "folklorist" reading of jokes, insisting instead that "a joke is a form of social communication" and as such should be viewed "in relation to their communicative context" (p.269-270). That is to say that rather than discrete utterances that occur in the abstract, Billig believes jokes and their effects are bound up with the context in which they are performed. He points to the "meta-discourse of humour" (p.270) to expand on this, noting that jokes do not just 'occur' but are introduced and framed in certain ways that identify them as jokes. The meta-discourse around jokes then can provide as much rhetorical power as the jokes themselves, working to normalise and justify the use of unacceptable words or stereotypes in certain contexts, even when the content of the joke itself may not achieve a particular communicative aim. Billig bolsters this conception of a

context specific view of joking when he discusses Attardo's (2000: 793) view on irony as "relevant inappropriateness". Attardo, in his overview of the many definitions of irony, suggests that when a statement is made ironically or sarcastically, it is specifically the context in which the utterance is made that implicates its status as ironic. It is the paralinguistic cues which alert a reader to the joke nature of an utterance, and when many of those non-verbal aspects are removed through CMC, context becomes arguably more important for understanding the intent of a comedic statement.

When dealing with online situations where prosody and non-verbal cues like tone are removed, the framing of jokes *as* jokes becomes trickier. Not only do jokes need to be more explicitly framed as jokes to provide the correct meta-discursive context, but the vagueness of unaccompanied text provides an interpretive gap, where those who perform speech which is not received well by their audience may retroactively assign joke status. The misunderstandings caused by a deprivation of non-verbal cues can be harnessed and weaponised as a justification for poor speech.

Billig (2001) goes on to discuss the power of the "joke" label to allow for "openly proclaimed" racism instead of the "mitigated or denied racism that is such a feature of mainstream discourse on race" (p.275). Although racist jokes are proclaimed without any additional justification, Billig suggests that through the joke framing a speaker identifies themselves as not necessarily 'not racist', but not a "*real life* racist" (p.275). This implies that "real" racism is violent, extreme and practiced, in contrast with the abstracted joke racism. And although Billig's work focuses on the use of comedy to justify racism, similar concepts can be transplanted onto other aggressive forms of speech which are not necessarily policeable or as socially unacceptable, such as counter speech. While counter speech is a defined in its opposition to hateful speech, it can also present as combative and insulting to the individual. As with hate speakers, in deploying jokes and comedy, counter speakers can soften and mediate their aggressive or insulting speech, possibly in a way which reduces its stigmatising elements.

Other investigations into the rhetorical function of joking have found that even when used in a teasing or aggressive manner, jokes can have a community building and

reintegrative function. Terrion and Ashforth (2002) found that comedic put downs can have the ability to engender quick solidarity in temporary groups, while also having a disintegrative effect between the joker and their target. Coates (2007) discusses the collaborative and integrative nature of conversational humour. Coates' investigation focuses on the talk between friendly groups, as opposed to adversaries, but as is elaborated on in the definitions of hate and counter speech, these forms of speech do not only work to attack but also to convince and inform those who may be receptive. Coates (2007: 29) notes that in defining a "play frame" through joking, others are given an area in which to create and collaborate through linguistic techniques like co-constructions and repetition, the latter particularly relevant in this thesis' discussion of mimicry in counter speech. Haugh (2014, 2016) also notes the power of joking (or framing utterances as joking), in their ability to sign post "non-serious intent" (Haugh, 2016:120) and to use that non-seriousness to engage in "jocular mockery" (Haugh, 2014:76), where ridiculing can be achieved in a way which reduces the likelihood of aggressive retaliation. This ability to frame insult and ridicule as well intentioned or community building, whether to soften the blow to a target or to engender group membership beyond the target, is a useful tool in the application of both hate and counter speech.

### Politeness

Fraser (1990: 219) concisely summarizes four prevalent conceptualisations of politeness, discussing the "social-norm view, the conversational-maxim view, the face-saving view and the conversational-contract view", each having their own varied focus on historical-cultural perspectives, rational agency, interactional co-operation, face and institutional formality. While each of these perspectives have obvious merit, and may be touched upon in this chapter, the focus here will be on the face-saving view of politeness, propped up upon Goffman's (1971) notion of "face" and fully formulated by Brown and Levinson (1978).

Brown and Levinson (1978), in their theorization of politeness, note that in many types of verbal acts (requests, complaints, criticisms, offers etc.) there are levels of politeness that are deployed to counteract or smooth out the inconvenience or imposition that speech act makes upon the recipient. For Brown and Levinson, politeness is not about adhering to the

formality of the social interaction (although that may play a part), it is about acknowledging that speech acts such as these may in some way damage the "face" of either the speaker or the hearer. The severity of these face-saving strategies are changed to adapt to different factors, including familiarity between participants, level of inconvenience in the speech act, mode of communication or type of situation, among others. In adapting to these different influencing factors, politeness can take many forms from inclusive language and plural pronouns, to explicit apologies and indirect language containing what Grice (1975) termed 'conversational implicature'. Conversational Implicature is the inference embedded within indirect speech, provided upon the assumption that both participants are being co-operative and the success of which is gleaned from the coherence of the exchange as a whole.

Brown and Levinson provide options through which a request, or other imposing speech act, can be accomplished; going on or off record and going with or without redress. A speech act being on or off record depends on whether the intended outcome of that speech act is ambiguous or not. If an explicit request is made, or a criticism is stated out right, with no reasonable room for questioning the motive or intended consequence, an actor is deemed to be going "on record". Alternatively, when one goes "off record" ambiguous and indirect language is used that may be interpreted in more than one way, giving the speaker room to detach themselves from any interpreted intent. The example Brown and Levinson give is noting the difference between the phrases "can you lend me some money?" and "oh damn I'm out of cash!". The second phrase carries with it the implication that the speaker needs money, doesn't have any, and would like the hearer to provide some, but that implication depends on the hearer's interpretation and allows the speaker to feign surprise at an offer or back track if confronted about their intention. How polite each of these phrasings are depends on the factors mentioned before (familiarity, social setting), but going "off record" provides more space for saving the face of both the speaker and the hearer if interpretation were to go wrong or imposition were to be poorly received. However, this additional safety comes at the cost of possible negotiation or avoidance of the implied intent of the speech act.

Redressive actions are speech acts with attempt to "give face to" the addressee (Brown and Levinson, 1978: 69). Brown and Levinson note that redressing can be achieved through the application of either positive politeness (indicating that the speaker and hearers

wants are the same through in-group membership or affectionate evaluation) or negative politeness (addressing that the speaker understands and is apologetic of any imposition). So, while being on/off record deals with the specificity or vagueness with which a speech act is made, redress concerns itself with the strategies used by the speaker to soften and justify that speech act. The speaker can decide whether or not to redress on or off record speech acts, and which forms of redress to apply, each conferring different levels of "politeness". Brown and Levinson discuss further that going off record and with redress, what would be assumed to be a *more polite* way to construct a speech act, may not always cause the expected outcomes. Overly redressed off the record requests may expose themselves explicitly as the acts they are trying to hide, or worse, may alert the addressee that the request being made of them may be more inconvenient or more face threatening than they are:

"If an actor uses a strategy appropriate to a high risk for an FTA [face threatening action] of less risk, others will assume the FTA was greater than in fact it was, while it is S's intention to *minimize* rather than overestimate the threat to H's face. Hence in general no actor will use a strategy for an FTA that affords more opportunity for face-risk minimization than is actually required to retain H's cooperation. " (Brown and Levinson 1978: 74)

We can understand then that politeness strategies and their application is a much more subtle art in persuading and gaining co-operation than merely deploying as many face-saving mechanisms as possible in any given circumstance.

In the context of online hate speech, one could assume that there are at least three potential faces being delt with, the speaker, the target of the speech and the assumed audience. As noted in the discussions on hate and counter speech definitions above, the rhetorical quality of hate speech is not simply to do illocutionary harm to a target group, but to convince and persuade an audience that may or may not be reading. Since a majority of social media speech (particularly the data used in this study) occurs in an open form broadcast at least to the followers of the hate speaker or possibly to a wider user base, the hate speech is never a closed interaction between two individuals (Graham and Hardaker, 2017: 798). The "addressee" or "hearer" as labelled by Brown and Levinson (1978) includes both the target

and a possible bypassing 'public'. Where hate speech may be assumed to be impolite in its nature, politeness techniques are necessarily used to maintain the face of the speaker, the audience they are trying to convince and maybe even the target who they wish to demonise or degrade while providing themselves or the target with a linguistic "out" (Brown and Levinson 1978: 70) to avoid escalated confrontation.

## Impoliteness

As Graham and Hardaker (2017) note, research on impoliteness in microblogging sites and social media platforms is still relatively sparse but is growing steadily. Impoliteness, as with politeness, is not a static technique but is adaptable to and dependent upon the expected norms of an interaction. What may be face threatening and insulting in one situation may be dismissed as banter or friendly joking in another and so while it is a defined concept, its identification in situ is more difficult. Culpeper (2005) provides a useful definition of what impoliteness is and what it isn't. Beginning with what it isn't, Culpeper states that impoliteness is not "incidental" (p.36), "unintentional", "banter" or "bald on record politeness" (p.37). That is to say that impoliteness is not an accidental by-product, but the end in itself, and furthermore it is not in-group offensive joking or in-group enabled sternness. Culpeper summarises his definition of what impoliteness *is* when he says:

> "Impoliteness comes about when: (1) the speaker communicates face- attack intentionally, or (2) the hearer perceives and/or constructs behavior as intentionally face-attacking, or a combination of (1) and (2)." (p.38)

As Culpeper rightly notes, intentionality is a problematic idea to grapple with as it can only be inferred in communication. This idea is exacerbated further by Brown and Levinson in their suggestion that terming politeness "strategies" as such suggests a level of conscious awareness that a researcher or viewer may be projecting and that the rationale behind politeness or impoliteness may not be as agentic as assumed. Culpeper then suggests that there is a second 'layer' to impoliteness worthy of consideration: information within the offensive utterance provided with the intention to cause offense.

As with Brown and Levinson's breakdown of politeness strategies, Culpeper summarises his breakdown of impoliteness strategies:

> *"Bald on record impoliteness*: the FTA is performed in a direct, clear, unambiguous and concise way in circumstances where face is not irrelevant or minimized.
> *Positive impoliteness*: the use of strategies designed to damage the addressee's positive face wants, e. g., ignore the other, exclude the other from an activity, be disinterested, unconcerned, unsympathetic, use inappropriate identity markers, use obscure or secretive language, seek disagreement, use taboo words, call the other names.
> *Negative impoliteness*: the use of strategies designed to damage the addressee's negative face wants, e.g., frighten, condescend, scorn or ridicule, be contemptuous, do not treat the other seriously, belittle the other, invade the other's space (literally or metaphorically), explicitly associate the other with a negative aspect (personalize, use the pro- nouns "I" and "You"), put the other' s indebtedness on record.
> *Sarcasm or mock politeness*: the FTA is performed with the use of politeness strategies that are obviously insincere, and thus remain surface realisations.
> *Withhold politeness*: the absence of politeness work where it would be expected. For example, failing to thank somebody for a present may be taken as deliberate impoliteness." (2005: 41-42)

Recontextualising language from Brown and Levinson's strategies in some places, this breakdown provides a guideline for identifying instances where both direct and indirect forms of intentional impoliteness are used. Whether direct or indirect, the common link through each of these strategies is to attack the face of an addressee in a way that can be read as intentional by those receiving or observing the attack.

## SUMMARY

This chapter has provided an overview of literature relevant to the research focus of this thesis, understanding hate and counter speech as a phenomenon and the digital world,

particularly social media, as its site of enquiry. Beginning with a discussion of hate and counter speech themselves, this chapter synthesised definitions drawn from legal, online platform and academic sources (Williams, 2019, 2021; Chakraborti and Garland, 2012; Garland and Chakraborti, 2012; Richards and Calvert, 2000) to provide a solid base of knowledge about what does and does not constitute these concepts, from which to build the research. From here, this chapter moved to discuss the criminological theories most suited for the online digital arena and social media research more specifically. Garland's (2001) culture of control and updated applications of classicist criminology were explored for their fit into an expansive networked society where the boundaries between public and private are blurred and spatial and temporal context can be collapsed. This was then expanded upon in the next section, looking in-depth at the ways social media are designed, how interaction achieved and the impact those things have on the kinds of communication people engage in. Kiesler et al. (1984) provided a social-psychological explanation of how distance, anonymity and a-synchronicity in online communication have stifled and reduced the interactional feedback that structures face-to-face communication, and suggest ways it has been seen to negatively impact interaction. Finally, this chapter looked at the theory driving each of the analytic traditions deployed in this research. Looking at identity and shame (as proposed by Goffman (1967), Braithwaite (1989), Scheff (1987) and Lewis (1971)), Genre, constructionist genre and intertextuality (Swales, 1990; Briggs and Bauman, 1992), humour (Billig, 2001,2005; Terrion and Ashforth, 2002; Coates, 2007) and linguistic politeness and impoliteness (Brown and Levinson, 1978; Culpeper, 2005), this chapter has explored and engaged with the theoretical underpinnings of key discourse analytic traditions that most usefully fit with investigating hate speech and counter speech, with a particular mind towards its application in digital spaces, where interaction is texturally deprived at the same time as its impacts are amplified. This thesis will now turn to a discussion of the methods designed and employed in this study.

# CHAPTER 3: Methodology

## INTRODUCTION

This chapter will discuss the methodological undertakings and justification for the research presented in this thesis. The study undertaken in this thesis can be broadly categorised as an exercise in Computer-Mediated Discourse Analysis (CMDA) (Herring and Androutsopolous, 2015, Herring 2004). Informed by notions of symbolic interactionism (Mead, 2005), this thesis takes three distinct forms of discourse analytic tradition and applies them to different but related speech acts as they occur within the data. This chapter will begin with an overview of the CMDA tradition and a grounding of that tradition within the specific research carried out, followed by an explanation of the data used. Following that, the three specific areas that are covered in the analytic chapters of this thesis (identity, (im)politeness and generic intertextuality) will each be discussed in terms of their individual methodologies and theoretical underpinnings, followed by a final discussion of the ethical boundaries and implications of conducting fine grained qualitative analysis on digital social media data.

This thesis follows on from two MSc dissertations (Roach 2015, 2016), each exploring online hate speech, and later counter speech. Both of these pieces of research were limited by a more surface level engagement with the phenomena they were investigating. Each piece of work sought to code and count instances of speech which were either identified by algorithm or chosen without an in-depth understanding of what those instances of speech and their speakers aimed to achieve with them. In a progression from a quantitative analysis of online hate speech frequency to a thematic content analysis of hate and counter speech language features, the methods of this research were chosen to delve further into understanding the linguistic tools and strategies employed by both hate and counter speakers as they occur *in situ*. As social media becomes all but universal in modern life, the discourses that shape our everyday interactions online require deeper and more fine-grained analysis to understand how one goes about *doing* being online, particularly when it comes to possibly violent and offensive interaction like hate and counter speech. While engaging with symbolic interactionism and CMDA generally, this research is still couched within a criminological

framework, utilising these methodological and analytic tools to understand how criminal (and sub-criminal but harmful) online behaviours are accomplished, how those accomplishment mechanisms seek to avoid penal and social/situational regulatory mechanisms, and how the informal social policing of these discourses is achieved.

Moving forward, this chapter will begin in earnest with a discussion of the primary methodological drivers that informed this research, before identifying the specific analytic tools within those traditions which powered each different form of analysis

## Computer Mediated Discourse Analysis, Symbolic Interactionism and Meaning making in online data

The overarching methodological approach of this research was chosen after implementing different complimentary forms within the discourse analytic tradition upon the collected data set. Herring (2004) herself suggests that Computer Mediated Discourse Analysis is in fact "best considered an approach, rather than a "theory" or a single "method"" and that CMDA provides a "methodological toolkit and a set of theoretical lenses through which to make observations and interpret the results of empirical analysis" (p.342). Indeed, through a semi-inductive process of emersion within the data relevant methodological frameworks emerged which were then retrospectively understood as collected under the umbrella term of CMDA. Again, in her breakdown of CMDA as a research approach she illustrates this bottom-up methodology when she states that "most CMDA research does not take as its point of departure a paradigm, but rather observations about online behaviour as manifested through the discourse" (p.358). This then legitimised and gave license to the research to draw "from discourse analysis and other language-related paradigms" (P.357) as were relevant and necessary to investigate the phenomena as discovered. This is not to say that the analytic frameworks found within this thesis were thrown together haphazardly, but that in each area of inquiry (Identity, (Im)politeness and Generic Intertextuality) the analysis could be rigorously structured by the paradigm most appropriate. Garcés-Conejos Blitvich and Bou-Franch (2019:3) note as much in their discussion of digital discourse tools when they

state that “some of these methods and tools may need to be critically assessed and reflectively adapted, and perhaps also expanded and even combined with others to suitably account for the communicative practices that occur in the digital world”.

## Reading Digital Data for Discourse Analysis

In their discussion of digital practice, Jones, Chik and Hafner (2015) summarise discourse analysis as a frame of enquiry, and digital text as a specific setting for that enquiry. They note that, like Herring, they approach discourse as a means to “build and manage… social worlds” (p.3) and that their interest is in what “situated social practices…people use discourse to perform” (p.2). So too this thesis is focused on what is done with language to create and maintain interaction, self and society. Jones, Chik and Hafner introduce the idea of the “texture” (p.5) of kinds of text, and the unique qualities the digital space brings to discourse. While not the only important texture, what is considered key during this research is that which is stripped away when comparing online with face-to-face interaction. As both a theoretical driver and a methodological convenience, the data used throughout this thesis are treated as bare text only. This is how the data is presented after being collected through the API scrapping tool COSMOS. COSMOS is a “distributed digital social research platform, providing on-demand analytics for the purposes of observing and inductively interpreting socially significant evidence gathered via the emerging uptake of social computing” (Burnap et al, 2014) and was used to collect and collate relevant data from ‘sentinel’ social media accounts. In processing the data, this tool presents a populated spreadsheet containing usernames, handles and the content of their posts in a linear temporal fashion, but does not provide any demographic information, images or emoji, and as such the data necessitated treatment as bare text.

Additionally, it was important for this thesis to treat the data in this fashion, because unless explicitly stated by a user, one could not assume prior knowledge or familiarity between users. This became a key epistemological principle of the research, that what was being interpreted in this analysis was taken only from what was being done in text by the interactants.

The key defining "texture" of the discourse from the perspective of this research, was that it was abstracted from demography, geography and the bounds of the classical temporality afforded by other synchronous forms of discourse. Creating, maintaining and influencing the social and the self through text only emphasises the impact and power of speech as action in this setting.

## The Methods of Symbolic Interactionism

This chapter will now move to another of the main methodological drivers of this thesis, symbolic interactionism. While this theoretical and methodological lens is most explicitly explored in the chapters on identity (chapters 5&8), its influence is felt throughout this research and its choice of analytic focus. Drawing primarily on the foundational work of Mead (1934) and Blumer (1969), this research takes as a key methodological stance that, particularly exacerbated within the restrictive settings of online social media, it is the interaction between speakers, audiences and generalised others which creates, develops and maintains the self and society. In his seminal work, Mead provides a description of symbolic interactionism which is seemingly idealised and typified in the text and language focused interactional setting of online social media. Mead (1934) suggests that:

> "The language process is essential for the development of the self. The self has a character which is different from that of the physiological organism proper. The self is something which has a development [and] arises in the process of social experience and activity" (p.135).

The emphasis on language as the tool for the development of the self and the understanding that the self is bound up and kept within that language, separate from the bodily organism, reinforce the importance of understanding the digital self as a site of real interaction, real identity and real harm.

Furthermore, Mead expounds upon the legitimacy of the pluralities of the self in different settings and contexts, as well as the power to develop the self in relation to the self

through the social intercourse of addressing the self and other simultaneously through writing, noting that:

> "We divide ourselves up in all sorts of different selves with reference to our acquaintances. We discuss politics with one and religion with another. There are all sorts of different selves answering to all sorts of different social reactions." (p.142).

In his example he cites the writing of a book as the medium for this interaction, but there is a clear parallel to the publishing of a social media blog or status update where the self is asserted and maintained through language that addresses both the writer and a generalised other or audience when no specific one is chosen. When a piece of social media content is created, be it Tweet, YouTube video or Facebook status, the recipient audience for that may be no one or it may be everyone possible, but the symbolic interaction of creating and recreating the self is evident regardless. This philosophy is an important methodological underpinning of this research, that what occurs in the data collected are interactions that can be taken and read in themselves as functional and illocutionary. What happens in the interaction, whether identity creation and maintenance, polite or impolite violent speech or the use of generic intertextuality to deploy or counter hatred, is readable, researchable, meaningful and understandable as legitimate within that interaction.

### *Meaning making in Computer Mediated Discourse Analysis*

The legitimacy of meaning making is reconstructed specifically for the Computer Mediated Discourse (CMD) arena by Herring and Androutsopoulos (2015) when they note that:

> "In CMD, meaning is constituted and negotiated almost entirely through verbal discourse. This is especially true in textual CMD, in which context cues are reduced relative to face-to-face communication" (p.133).

Herring and Androutsopoulos go on to list ways in which meaning is created in textual discourse, expanding particularly on the blatancy of that meaning making in "performative

utterances" that "do by saying" (2015: 135). They provide the example of judicial and legal language ("I hereby declare") used to draw on the indexicality and intertextuality of the authority bound with the associated speech genres they invoke. The power of the meaning created by those performative utterances is emphasised and magnified by the restrictive settings of CMD.

With a methodological emphasis on the power of meaning making in textual utterance, Herring's (2004) outline of the CMDA approach invokes Goffman (1959) to provide an explanation for the efficacy of researcher observation as a legitimate interpretation of that meaning. Herring (2004: 342) notes that a core assumption, and by extension site for analysis, for CMDA is that "discourse exhibits recurrent patterns" and that they may be produced "consciously or unconsciously (Goffman, 1959)". In identifying the possibility that active awareness of used strategies may not always be the case, Herring argues that the observations taken by analysts may be more reliable and generalizable than reports gained through interviews with the participant. What matters, CMDA may argue, is that which is seen to occur in the interaction, not necessarily the unknowable authorial intention that leads to an utterances production. This links back to methodological conventions of conversation analysis (one of the many tools afforded in the CMDA toolkit) and Sack's ethnomethodological understanding of meaning in action (Mair and Sharrock, 2021)

With an overview and justification of the philosophical, epistemological and ontological underpinnings of the broad research approach that will be employed in this research, this chapter will now turn to the specifics of its execution.

## The Nuts and Bolts of Digital Data

### What are the data and how are they treated?

For this study, the collection was performed on two 'sentinel' accounts. These are dedicated social media accounts used to re-publish and broadcast hateful utterances as a form of shaming and non-institutional policing. The ethical drivers that influenced this

decision will be expanded upon further in their own section, but the choice to collect from these sentinel accounts allowed the research to be focused in on the particular area of interest (hate speech and its response) without the additional labour of combing through the, on average, 500 million tweets per day[5] to identify the instances of interaction. The sentinel sites used for this data collection focused on racist and homophobic hate speech. The data they broadcast is found either through searching offensive or stereotypical words and phrases associated with their respective form of hate, or by having that content reported to them through 'tagging'.

While racist and homophobic hate speech are by no means the only forms of hate speech found online, the choice to focus on these forms of hate speech was both pragmatic and methodological. Firstly, these forms of hate speech are two of the most prominent, both historically in their occurrences and in their academic investigation. This meant, rather unfortunately, that there was (and is) a wealth of examples online of this form of hate, and along with those occurrences, instances of counter speech found in response. Additionally, with the academic literature detailing the exploration of these two forms of hate being so proliferated, there was a great deal of existing resources available to draw on to understand the design of both the hate and its response. And while all hate is not alike in its target, its motivation or its production, there are important similarities across their formulations. Some of the similarities between the occurrences of racist and homophobic hate speech are illustrated in this thesis, and it is hoped that the understanding of those similarities (and differences) may be extrapolated to aid in the understanding of other forms of hate, such as that experienced by the transgender community or the disabled community who themselves are often targeted in online settings.

Herring (2004: 350)) grapples with the pragmatics of this style of data collection, noting that in CMDA typically, data is "logged or culled from online archives" and not generated in an experimental fashion, to provide 'natural' data that is practical for discourse analysis. She also discusses how in reducing the randomness of the sample by necessarily paring down the mass of available data this snaps into focus the context necessary for

[5] https://www.dsayce.com/social-media/tweets-day/

interpreting discourse analysis results. In sampling through sentinel sites geared specifically for collating hate speech, this analysis is particularly focused on offending speech and its response without having to parse a great deal of extraneous data that may provide a grand global context but does not aid the aims of this research. Herring (2004: 351) developed a table of sampling techniques commonly used in CMDA and concluded that the advantages offered by creating your sample by theme and time have made them the favoured choices for this kind of research:

TABLE 12.2. *CMDA Data Sampling Techniques*

| | **Advantages** | **Disadvantages** |
|---|---|---|
| Random (e.g., each message selected or not by a coin toss) | Representativeness; generalizability | Loss of context and coherence; requires complete data set to draw from |
| By Theme (e.g., all messages in a particular thread) | Topical coherence; a data set free of extraneous messages | Excludes other activities that occur at the same time |
| By Time (e.g., all messages in a particular time interval | Rich in context; necessary for longitudinal analysis | May truncate interactions, and/or result in very large samples |
| By Phenomenon (e.g., only instances of joking; conflict negotiation) | Enables in-depth analysis of the phenomenon (useful when phenomenon is rare) | Loss of context; no conclusions possible re: distribution |
| By Individual or Group (all messages posted by an individual or members of a demographic group, e.g., women, students) | Enables focus on individual or group (useful for comparing across individuals or groups) | Loss of context (especially temporal sequence relations); no conclusions possible re: interaction |
| Convenience (whatever data are available) | Convenience | Unsystematic; sample may not be best suited to the purposes of the study |

(Fig.3 - Herring 2004: 351)

The data for this research was chosen and collected by theme, that theme being hateful speech and its responses. By utilising sentinel accounts as a lightning rod for data

collection, the categorisation and distinction of those themes were created by an outside, visible public source. This removed the labour of defining the data collected as being hateful or not from the researcher, as that had already been established by the operators of the accounts. As will be discussed below, the work of Sacks' (1972) on membership category analysis, particularly the "viewer's maxim" provides unique justification for the reliance on those doing the category work of identifying hate speech. Reynolds (2017:100) discusses the viewer's maxim, quoting it as:

> ''if a member sees a category bound activity being done, then, if one can see it being done by a member of a category to which the activity is bound, then [sic] see it that way''

By this, one can understand that if a member of a category identifies an activity as being bound to that category, then one should see it that way. The sentinel accounts can readily be identified as producing or facilitating counter speech against hate speakers, and if they as members of the category 'counter speakers' are engaging in the category bound activity of opposing hate speech, and other members (or counter speakers) can see it being done, then so should the analyst. This technique for identifying hateful speech, as with computer aided identification, leaves itself open to misinterpretation or the misreading of irony or sarcasm (the optionality of a sarcastic reading itself is a pre-emptive protective technique that may be deployed by hate speakers, and is discussed in the analysis of this thesis). Moreover, in relying on unknown users piloting sentinel accounts, this research cannot know the user's demography, biases or techniques used to identify hate, which is an unfortunate shortcoming of this research.

Additionally, in utilising the definitions of counter speech discussed in the generic intertextuality chapters of this thesis (chapters 4&7), one could assume all opposing responses collected in the ensuing thread underneath the established hate speech that did not agree or confirm the hate speech, could justifiably be considered counter speech. As Herring notes, by collecting the data in accordance with these themes it ensures "Topical coherence" and a dataset which is "free of extraneous messages". By relying on the coding of speech as hateful by the accounts themselves, instead of the COSMOS sentiment analysis or

other machine classification tool, this research was able to circumvent the methodological problem of relying on fallible computer-generated algorithm to identify hate (which are routinely found to generate false positives and negatives). However, while this does avoid the issues with computer generated hate analysis, it does not avoid human interpretation error, bias, or the flaws that may occur in their identification methods (Twitter search tools, reports from other human users).

This also brought into light the contemporary and shifting nature of language and its understanding over time. What was engaged with here is a snapshot of language use, judged at the time by public members to be hateful. The efficacy of the initial analysis performed by the account holder is up for debate and is often the focus of the ensuing thread beneath, but again the onus is not upon this research to decipher what is or is not hateful, but to identify and understand the communicative action used by social media users to perform their linguistic actions as the occur. As is shown in the analysis chapters, much if not all of the data presented can be identified as hateful when compared against the definitions provided by the platforms, but importantly for this research, this data has already been identified as hateful by a sentinel site and exposed as such. Whether the operators of those sites are making their analysis in concert with the accepted definitions or not (and their bias may well influence their decision to publish), the utterances analysed have been broadcast as hateful and are under discussion by the digital "public".

## Data Collection and Presentation

By using the COMSOS collection tool 100 threads were scrapped from the sentinel accounts @YesYoureRacist and @Homophobes starting in late 2014 and moving backwards in time until the 100-thread threshold was met, ending in 2013. While these are historical data, particularly in the context of the accelerated movement of online life and technology, this data set is not without the possibility of novel and contemporary findings. Social media and the internet have no doubt developed greatly since the collection of this data, especially when taking international political upheavals such as Brexit and the presidency of Donald Trump into account. However, this thesis would argue there is a great deal of value in understanding the construction and deployment of hate speech and counter speech at this

time as both a foundational piece of work upon which to build, as well as an important and transplantable piece of knowledge in itself, that has important implications to our understanding of hate and counter speech in the present day.

Each of the 200 threads collected contained at least one response, but in some cases more than 140, so the totality of the utterances that were parsed through in initial engagement was considerably higher. There was a total of 1,708 tweets collected in the racist hate speech data set and 1,849 tweets in the homophobic hate speech data set. While the @YesYoureRacist and @Homophobes accounts deal largely with racism and homophobia respectively, they also amplify instances of religious and xenophobic hate, and LGBTQIA+ hate more generally as well. The specificities of each of these individual forms of hate speech are no doubt worth in-depth analysis in their own right, however the aim of this research is to explore the linguistic and communitive properties of 'hate speech' and 'counter speech' as genres in themselves. It is for this reason that the individual 'flavours' of hateful language found within the analysis are presented alongside one another, with little reference to any distinguishing elements.

The sample of hate and counter speech presented in this thesis were chosen particularly for the purpose of illustrating and illuminating the linguistic techniques that were derived from this analysis. Any individual utterance with have a plethora of possible linguistic interpretations, and while each of the linguistic strategies discussed here may not occur in every thread collected, the sample presented are representative of the wide majority of hate and counter speech collected, where in techniques like legitimation, oppositional identity, comedy, mimicry and (im)politeness are frequently found.

The data collected and presented contains no identifying demographic information and as such is not broken down by gender, age, location or other. For both convenience and epistemological reasons this caused no great issue but is still worth acknowledging. In parsing the data there is no doubt that some users took the initiative to investigate the profiles of other speakers and made use of information from profile pictures and biographies to inform their speech. Multimodality (Thurlow, 2017) and the inclusion of other forms of media such as image, video or emoji in CMDA is an important emergent discussion, however because of

the format of the collected data it is not included in this research. Additionally, from an epistemological and methodological standpoint, because these data were retroactively collected and direct contact with the users was unavailable, it was never possible to entirely confirm if users were informing their speech with demographic data gleaned from a profile picture, biography, previous familiarity or any other means. Moreover, the veracity of any information taken from pictures or account biographies could not be confirmed to be accurate in themselves, so this redoubled the focus of this research to be concerned with what was done with language in situ. The identity maintenance, illocutionary action, community and world building observed and reported were all taken as symbolic action within the language used.

As noted previously, the data analysed for this research is taken from online social media platforms, specifically from "sentinel" profiles within those sites. The decision to use these profiles as data collection tools provided a host of pragmatic and logistic benefits, as well as some methodological draw backs. One particular issue with the use of these sites is lack of transparency surrounding the practices of those in control of those profiles. They are private profiles who do not make explicit their collection process (how they find and decide which utterances to broadcast) or their own standards for what constitute hate speech. However, the use of these sites removes the onus of responsibility from this research of attributing hate speech labels to the data collected, an endeavour which is outside of the scope of this research. This thesis does not seek to implement its own judgement upon hate and counter speakers (an action which in itself might generate more harm for online users) but to discuss how interactants perform their speech genres when they have been identified as such in the wild.

Moreover, an alternative route such as analysing privately collected data from a source like Tell MaMa (2019) would run the risk of exposing un-broadcasted hate speech to a wider audience and incurring harm. Additionally, the dissemination of utterances that may be designed for one audience to a broader viewership[6] has already taken place. Using publicly

[6] Context Collapse (Marwick and boyd, 2010), discussed further in Chapter 5

available data such as this significantly reduces the likelihood of reproducing the data in a context that may cause more harm to both hate and counter speakers.

It is also important to address the a-synchronicity of the data, which is not so much an issue as it is a feature of computer mediated discourse and one that may be exacerbated by the use of sentinel accounts as well as the retroactive collection method. In their discussion of "interaction management", Herring and Androutsopolous (2015) suggest that "perhaps more than any other aspect of CMD, interaction management is shaped by the medium characteristics of CMC systems" (p.136). That is to say that when addressing interactional data, the specific characteristics of computer-mediated-communication systems, especially those which lend themselves to a-synchronicity and disruption, are particularly prevalent in their influence over those interactions. They go on to say that "in environments such as multiparticipant chat rooms,…", social media platforms providing practically limitless possible participants, "…where disrupted adjacency is common, communicators may even come to accept tenuous or "loose" relatedness of adjacent turns as normal" (p.137). In this instance the a-synchronicity and disruption are compounded by the sentinel account broadcasting the initial utterance at a delay to a much wider audience. Additionally, there is the standard platform characteristic that threads are not produced in an entirely linear fashion, with responses to the original tweet coming in parallel and those responses creating diverging threads from the same origin point that may or may not be being generated simultaneously.

The data as presented in the populated spreadsheet appears linear and temporally organised, but the appearance of 'handles' indicating to which individual or collection of individuals the response is aimed show this not to be the case. To avoid confusion and complication by concurrent threads, intersecting and interrupting one another, individual utterances, adjacency pairs and threads are pulled from the maelstrom of data to illuminate and focus the examples. In certain cases, the same instigating hate speech utterance was presented with a different response thread following it. In these instances, the example shows a specific interactional thread of linguistic turns as they occur between those two (or more) participants with the additional unrelated white noise removed. This is all employed as a means to deal with the a-synchronous and messy nature of social media data that necessarily involves so many concurrent participants. The data was then treated as a-

synchronous, but presented in a linear fashion to illustrate the adjacent interactions as they would occur, were they to be viewed as a conversation in a more traditional sense.

It is worth noting also that on occasion throughout this thesis threads may be used multiple times in different analyses. This decision was not only taken because those threads were the most exemplary illustration for each individual strategy discussed, but also to illuminate the notion that strategy and genre are not deployed exclusively or with total accuracy. Language and interaction are co-constructed and rely on audience interpretation guided by context (Jones, Chik and Hafner 2015: 66), generic precedent (Briggs and Bauman 1992: 163) and the indexicality of online identity work (Bucholtz and Hall, 2005: 953) all of which can occur alongside and in tandem with one another. As such each of these threads can be seen to contain a multitude of linguistic strategies, each of which can be interpreted differently dependent upon the audience of that message.

This chapter will now turn its focus to the discussion of the three specific analytic methods implemented in each of the analysis chapters of this thesis. In each instance the specific method of discourse analysis will be examined, with a focus on its re-specification for online analysis and CMDA. In understanding, designing and utilising each analytical style, attention is paid to the theory and method of the original offline variants, which are then compared to their development in digital spaces and any unique affordances that were made in this research to properly adapt to the style of data and unique discourses that were used.

## The Analytic Tools in the CMDA Toolbox

### Genre and Intertextuality

In extensively exploring the theory of genre as it relates to hate and counter speech and problematizing that form of linguistic categorisation in the literature review, much of the important methodological traditions that define this style of enquiry are necessarily covered already. It is for this reason that, to avoid repetition the discussion in this section will be more

particularly focused (as will the explanation of each of the other methods) around the justification of deploying this method, given the research area of hate and counter speech.

The initial foray into applying genre analysis to online hate and counter speech began with a basic understanding of Swales' (1990) formulation of Genre analysis and sought to identify and discuss the steps and moves (Swales 2004, Cotos, Huffman and Link 2015) that made up those genres. It was taken as granted at the time that hate and counter speech were viewed as specific speech genres and the intention was to illustrate a typology of moves or "rhetorical unit[s] that perform[s] a coherent communicative function" (Swales, 2004:228-229). Hate and counter speech appeared, to the researcher at least, to be visible and obviously categorizable forms of language that could be coded and counted. This may be the case to some extent, but in investigating the literature following this assumption it became clear that neither of these forms of speech had been explicitly addressed under the framework of genre. Moreover, the creation of a comprehensive, let alone exhaustive, typology of the moves that categorise these genres would require a much larger dataset and a much more quantitatively focused analysis than was preferred for this thesis.

In extended exploration of the literature around genre study and a continued immersion within the data, it became apparent that the use of genre was far more constructionist and malleable than the constraints of genre as a framing tool suggested. In noting that counter speakers were able to identify hate speech when it was not explicit according to the proposed definitions (and that hate speakers understood their speech to be hateful through their pre-emptive justifications), it suggested that there was a collection of linguistic strategies that were used to invoke their chosen speech genres in a way that was identifiable to readers but still maintained a safe distance from them. This aligned with writing by Briggs and Bauman (1992) on intertextuality within and between genres, which was applied specifically in online spaces by the likes of Zidjaly (2010) and Vásquez (2015). Zidjaly notes in her discussion of intertextuality and digital identity that:

> "Intertextual reshaping of texts has a wide variety of interactional functions, including building shared communities (Becker, 1994), accomplishing tasks (Tovares, 2005), creating involvement (Tannen, 2007), and constructing subtle layers of meaning

(Gordon, 2009). Intertextuality, moreover, has been analytically linked to identity construction." (p.193)

This summation neatly covers a range of the key accomplishments made by hate speakers in their attempt to perform their chosen speech genre without explicating in a way that would identify them unequivocally as hateful, or in ways that can provide an intertextual gap through which to avoid condemnation. Similarly, counter speakers were seen often to engage in combative speech which accomplished the "task" of opposing hate speech but by using techniques not expected or typified in previous definitional work on the subject.

With this all-in mind it appeared important and useful to employ Swales' (1990:46) criterion to categorise hate and counter speech as genres of language in themselves and investigate the data to understand how social media users deploy the tropes and features of those genres *and* how they use intertextual knowledge of offline and online hate and counter speech to achieve the aims of those genres in new and novel ways. This was particularly inspired by the recurring theme, both in the data and in general engagement with social media, that people still do actively and routinely achieve the communicative aims of each of those genres while under the constraints of the platform and contemporary society.

### Ethnomethodology, Conversation Analysis, Membership Categorisation Analysis and Identity

The identity focused analyses of hate and counter speech found in this thesis are based around the core principles of EMCA (ethnomethodology/conversation analysis), in particular the membership categorisation analysis as developed by Garfinkle (1967), Goffman (1959), Sacks (Sacks and Jefferson, 1995) and Schegloff (2006). This is also the analytic framework most explicitly informed by Mead (1934) and Blumer's (1969) Symbolic Interactionism, although their influence can be felt throughout the entirety of this thesis. As with each of the discourse analytic variants employed in this research, this methodological framework was decided upon by semi-inductive means. After prolonged engagement with the data, the recurrent theme of identity and its communicative use became apparent. In grappling with identity as a site for discursive analysis, the restrictions and unique facets of

social media and the data formatting made clear the importance and the in-situ power of identity as a tool for speech action.

Pioneered by Garfinkle (1967) and Goffman (1959), and later developed by Sacks (1995) and Schegloff (2006), this form of analysis takes as a central tenant that "conversation is not 'just talk'; it achieves *social actions*" (Benwell and Stokoe 2006:59). Jaworski and Coupland (1999) collect a range of established definitions of "Discourse Analysis" more generally and note a common theme throughout in the idea of "Language in use" (p.3). They suggest that regardless of the specialised tradition chosen within discourse analysis, language and text do not sit as abstracts, but are achieving action. Jaworski and Coupland expand upon this further, highlighting the idea that discourse is more than mere language in use, but takes on reflective and critical aspects, in that it is language in use *relative* to "social, political and cultural formations" which shapes, and is in turn shaped by those formations. In investigating racist and homophobic hate speech throughout this thesis, there will inevitably be some necessary discussion of their impact on the wider world. While the main analytic focus of this analysis is on the discursive traditions mentioned above, it would be an impossibility to ignore entirely their political and cultural influence. These more critical variations of Discourse Analysis, as championed by van Leeuwen (1993, 2007) and Reisigl and Wodak (2001, 2016), are also deployed in this analysis to work in concert with the EMCA and MCA traditions driving the heart of this discussion. Combining these analytic traditions under the Computer Mediated Discourse Analysis (Herring, 2004) banner will help to tie the "conversation as social action" analysed to the social and political surroundings in which it is found.

As noted, explicit references to the social media platform from which these data were retrieved are omitted. Pragmatically and methodologically, this situates the research within the bounds of the agreed upon ethical standards, while also focussing further on what is specifically happening within the text, as preferred within EMCA traditions. While the descriptive "physical" markers found in profile information (such as visible race, gender, sexuality etc) may serve as hateful triggers, within the data collected hate speakers are exclusively found referring explicitly towards their targeted *group* or a non-present individual, not a specific member within the interaction. For example, hate speakers have been found to identify "black people", "gay people" or public figures, but never a civilian with whom they

are interacting. Because they are not attacking or responding to a particular person based on traits they may have gleaned from their profile pictures or account names, the addition of those multimodal expansions on digital discourse analysis which are being spearheaded elsewhere (Garcés-Conejos Blitvich and Bou-Franch, 2019) were foregone within this analysis. Moreover, Sifianou and Bella (2019) neatly summarise the argument for addressing the content of social media data as a key analytic driver in opposition of a more multimodal approach when they note that:

> "Posters' potential anonymity and the disembodied nature of identity (not connected to a physical appearance or non-verbal behavior) along with the fact that Twitter user profiles are very brief (maximum 140 characters) prompts followers to draw conclusions primarily on the basis of the content of tweets. " (p.345)

While this research does not explicate the specific platform, the encouragement of brevity and disembodied anonymity is a recurring feature among many of the most highly populated social media sites. Identity work, as will be illustrated in the analysis, can be done, and is done, within the text of the discourse, and for this analysis does not require the additional demographic information that may be found in an accompanying profile to do so. Benwell and Stokoe (2006: 63) centre their research on "questioning *how* conversational actions are accomplished as the systematic products of sequentially ordered interaction rather than *why* they are performed", a sentiment that is followed here. While the contextual restrictions enforced or encouraged by social media platforms will affect the form that the interactions take, it is the focus of this research to understand *how* utterances accomplish their communicative act and not predict or presume the intent of why they were performed.

That is not to suggest, however, that the anonymity provided by the tenets of CMC and this analysis' lack of interest in demographic factors is irrelevant to this analysis. The level of anonymity provided in these situations is itself a defining driver into this investigation. In their discussion of the Social Psychology of CMC, Kiesler et al. (1984: 1125) suggest that the "absence of regulating feedback" and "depersonalization" of CMC provides a specific "dramaturgical weakness" to interaction. However, implementing a symbolic interactionist frame of constructed, interactive identity, this dearth of demography allows an analyst the

ability not only to discern what identity work is being done with language, (assuming a lack of participant influence from the demography), but more accurately shows how identity work is done in the eyes of the interactants, what ethnomethodologists would term the "members' methods" (Carlin, 2021). In general, it can be assumed that all first-time interlocutors are privy to the same amount of demographic information as the analyst, and so are required to derive and create their sense of identity and oppositional identity through language in use.

Velody and Williams (1998) use the term "realist residue" to describe known identity categories that are unintentionally deployed in analysis by researchers. A known gender, race or sexuality may well influence a researcher to apply those identities or presumed facets of those identities, in their analysis rather than uncover their constructions within the data. In removing the small amount of demographic data that can accompany the text of social media data, the researcher can hope to minimise that residue. However, it should be noted that because of the category groups that are being discussed (races, religions, sexualities etc) the influence of realist residue, even if actively avoided, may arise based on the prejudged assumptions about certain actors (for example, if anti-black hate speech is observed, a researcher may presume that speaker to be white, or at the very least not black).

Furthermore, these methods are informed by Wooffitt's (2005) exploration of conversation analysis and discourse analysis, particularly the rhetoric's of authority and persuasion. He notes that discourse analysis as a tradition began with the problematising of scientific or factual language as taken for granted and this rang familiar in the recurrent use of authoritative identity creation by hate and counter speakers found in the data. Interestingly and alternatively, both hate and counter speakers were also seen frequently to invoke othering identities as a means to bolster their own legitimacy. Wooffitt seems to tackle these ideas head on when he states:

> "Our everyday life is marked by minor dispute and disagreement: mundane interaction is rife with arguments, accusations, rebuttals, blamings, criticisms, complaints and justifications. And there are occasions in our lives when we make more contentious claims, for example, accusing someone of inappropriate behaviour, or reporting unusual experiences. It is possible that these kinds of discursive action will

> receive unsympathetic, sceptical or, indeed, hostile responses. Here again we can ask: what are the resources through which controversial reports are constructed so as to appear reasonable and robust, and to anticipate sceptical responses?" (p.93)

The final sentences in Wooffitt's statement acted as a basis for focussing the analysis, respecifying general "resources" as specific forms of identity work for constructing "controversial" hate and counter speech. It also provided a neat framing for the occurrence of hate and counter speech online in terms of mundanity and contentiousness. While hate speech is seen to be a deviant and policeable form of language, it is often treated as mundane, particularly by those attempting to *do* and to normalise that form of speech through authoritative identity. Similarly, counter speech is often violent, offensive and combative in its deployment, but treated as a matter of course by those who may find themselves in online 'public' spaces.

Bucholtz and Hall's (2005) influence is also felt in the analytic practice of this research, taking inspiration from their definition of identity as "the social positioning of self and other" (p.586), their view of identity as emergent in interaction and their inclusion of "indexicality" (p.587) as a necessary strategy for defining those self/other binaries. Those indexicallised binary categories that are often invoked and weaponised in antagonistic interaction, particularly those based specifically around the antagonism of group membership, inter-group hierarchy and violence, were shown to be an important and frequent occurrence. While Bucholtz and Hall suggestion that "it is the constant iteration of such practices that cumulatively produce not only each individual's gender, but gender itself as a socially meaningful system" (p.590) provide a framework that acutely fit onto the structure of hate and counter speech, where identities used, deployed and compared reinforced the systems and binaries that were being summoned. Black/white, gay/straight, local and foreign were all identities being used, but simultaneously systems of categorisation that were being ratified and authorised by their use.

The final theoretical driver for this analysis is Sack's membership categorisation analysis (MCA), as discussed by Silverman (1998). Mirroring Bucholtz and Hall's ideas of identity in interaction, MCA describes the use of categories and indexical references to

categories to provide identity for, or give identity to, members in interaction. In his discussion of "inference rich categories" Silverman (1998: 74) explains that, when invoked, categorisation by gender, race, political affiliation or nationality (among other things) can infer a great number of associated or accompanying characteristics. The power of those inferences is often overwhelming and can do an incredible amount of rhetorical work without the additional labour of explanation. When considered in online spaces where brevity is prioritised and the ephemeral nature of identity becomes magnified, reliance on category and indexical work to create and maintain the self (and the other in opposition) becomes a useful tool. In concert with Bucholtz and Hall above, the power of these categorical invocations provides not only for the construction of identity, but for the solidifying of the binaries and systems that these categories lie within.

Some examples of the existing literature on digital identities (Herring and Androutsopoulos, 2015; Benwell and Stokoe, 2006) are bound up with its connection to community and community building online. Social media identity work is often framed as relating to the specific community it finds itself within; social media sites are built around friends, family and followers, blogs, sub-reddits and YouTube channels often follow a topic, etc. Twitter is similar also, except with a lighter touch in terms of defining and maintaining group membership since the main tool used for that group building is uni-directional "following". So posts, utterances and the identities created and maintained within those are not always exposed to an expected group. The online phenomena of 'hate following' (following a profile you disagree with or dislike, with the aim to be informed when they post something you are opposed to) bypasses the common wisdom that your followers enjoy your content or are accepting of your performative identity work. Moreover, the use of sentinel sites, which cater specifically to engaging with and broadcasting content that they disagree with or dislike, transports messaging that was already displayed to an unconfirmed audience of "the generalised other" (Mead 1934:154) and expands that broadcast to an entire new set of audience members, many of whom find themselves predisposed to antagonism because of the base function of the sentinel.

The identity work done by hate speakers presumably does not account for the unknown addition of sentinel site audience members, but identity work they have attempted to achieve with their followers may be given new context and interpreted differently when

exposed to new spectators (Marwick and Boyd, 2010). The new context can change the efficacy and the effect of that identity work, and additional identity work done in response to a challenge will most assuredly be impacted by this change. Regardless of the change in audience, the uncertainty of audience in the initial utterance means identity through demography and familiarity will always be reduced, and the recreation and maintenance of identity is something that is seemingly taken up often by social media users of this fashion. The necessity of identity management and maintenance is emphasised even more so when the content being displayed is understood to be socially unacceptable. The presence of additional identity work in the form of authority claiming and othering may betray an understanding that hate speakers are aware from the get-go that their messaging is unacceptable.

To summarise, MCA provides a useful analytic tool for online hate and counter speech investigation, because so much of the identity work that is necessary to effectively produce these genres of speech are based around belonging, or crucially *not* belonging to a category of person. Membership to a particular category, be they majority, minority, in group or other, is something that is performed and weaponised in these combative interactions, and the analytic tool kit designed by Goffman, Garfinkle, Sacks and Schegloff provide a unique way of discerning the everyday methods interactants use to achieve those categories.

With an overview of Membership category analysis and it's fit into online identity work, this chapter will move on to the next analytic tradition deployed in this research, linguistic politeness and impoliteness.

## (Im)politeness

The analytic method for linguistic politeness, and impoliteness, used in the thesis stems from Brown and Levinson's (1978) seminal book on the topic, in which they lay out their theory and method for politeness as a "'ways of putting things'" that "are part of the very stuff that social relationships are made of" (p.55). In stating their initial method, Brown and Levinson admit to constructing a "Model Person" endowed with two special properties "rationality and face" (p.58). This model person is taken as their base assumption, which is

projected upon individuals whose politeness they try to understand and explain. The use of face here is inspired by the Goffmanian (1959) conception but is respecified in this context to include notions of positive and negative face wants, those being to be approved of by others and to be unimpeded by others, respectively. With this model person in mind, Brown and Levinson raise Grice's (1975) 'conversational implicature' to justify the assumptions made about inferring certain language use as indicative of a desire to manage face when it is presented. The analysis of this thesis takes this stance also, that those interactants found within the data are rational beings who are cognisant of their face (as well as the face of others) and intend to either protect, damage or make threat towards face by their use of politeness (or impoliteness) strategies.

In their discussion of politeness on Twitter, Sifianou and Bella (2019) provide an illuminating overview of the inherent murkiness of defining politeness as a *thing* to be studied, noting the necessary distinction between first and second order politeness, where one (first order) is defined by its explicit categorisation by lay people within the interaction and the other is its use as "theoretical concepts in a top-down model to refer to forms of social behavior" (Locher and Watts, 2005). Sifianou and Bella note than in more recent research there has been an indication that a complete distinction between the two views is difficult (if not impossible) but nod to the work of Ogiermann (2009) and Haugh (2010) in addressing that the theoretical approach to politeness may be valuable when dealing with naturally occurring data as has been afforded by the modern wave of logged internet data. It is for this reason that, although discussion of lay categorisation of politeness was not discounted, the main focus of this study followed politeness and impoliteness as a second order phenomenon, where the occurrence was identified and theorised without direct confirmation or explicit discussion by participants.

In discovering which politeness strategies were most frequent and pertinent to the topic, Brown and Levinson's typology (On/off record, redressive action, positive/negative politeness) was used as an analytic base, while also giving mind to examples from contemporary studies that bore their influence. Planchenault (2010) in particular, in her discussion of politeness in virtual transvestite communities, considered the use of "markers of solidarity" (p.94) in engendering group membership as a form of positive politeness. The

conciseness and repeatability of this strategy harkened back to Sifianou and Bella's (2019) summation on the necessity of brevity in online spaces as a constraint but also as a stimulus to creativity.

LeBlanc (2010), in her chapter on politeness and impoliteness in online community building mentions Hymes' (1962) idea of "communicative competence", and how this is a consequence of learning to communicate in an appropriate fashion for the setting and context they are in. Understanding the context of online social media and the lack of certainty of audience, as mentioned above, the influence of the "generalised other" comes to form an important part of creating that communicative competence and illuminates what can be assumed to be being achieved in the use of politeness in hateful speech. This then steered the framing of this analysis, to understand that politeness may not only be utilised as a face-saving mechanism for the identified target or the interactant pair, but as a means to reduce positive and negative face threat to the generalised other that the speaker means to build community with, through their othering. This settled the analytic lens as one that viewed politeness as a collection of strategies to maintain the self and face in the presence of face-threatening acts, but also to bolster community building through the same practises.

The analytic methodology for discussing impoliteness necessarily followed and built upon structure of politeness that preceded it. Culpeper's (1996) work creating an anatomy of impoliteness builds upon and respecifies Brown and Levinson's notions of politeness, inverting them and providing a useful framework through which the data was viewed. In the opening of his paper (2005) applying the study of impoliteness to TV show "The Weakest Link" he conveniently breaks down what impoliteness is *not* as a means to more readily define what is, and dismiss other impolite occurrences that are not specifically acts of impoliteness. His anatomies of what are and are not impoliteness helped frame the analysis for this thesis and sharpen the focus through which the data was viewed, specifically in terms of looking for instances of unambiguous and intentional impoliteness. Relating back to Herring and Androutsopolous (2015), understanding politeness and impoliteness, as with other forms of CMDA, is greatly influenced on the understanding of the medium characteristics found in online settings. This meant that prosodic features, such as intonation and volume, that may more readily indicate politeness and impoliteness, and differentiate themselves from

insincere forms, were restricted. This then needed to be considered when parsing statements and responses to ensure as best as could be that they were not being misinterpreted. Angouri and Tselegia (2010) address this specifically regarding online impoliteness when they, quoting Culpeper, suggests that:

> "a) these strategies are only a part of a wider repertoire participants adopt in such cases, since "impoliteness does not simply arise from any one particular strategy, but is highly dependent on context" (Culpeper et al. 2003: 1555) and b) the same strategies can be used for the exact opposite purposes (i.e., create and enhance rapport) depending on the overall context of the interaction." (p.66)

Angouri and Tselegia also mention "Non-standard spelling" (p.66) and "Emphatic capitalisation" (p.74) as ways of implying and accentuating impoliteness in particular, when accounting for the restrictions of text based online interaction.

Bousfield (2008) expands on Culpeper's work, creating a slightly altered functional definition for impoliteness, which forgoes the need for and understanding of intent from the hearer of the impoliteness. For the data used, this definition proved useful as often messages are left unanswered in social media interaction, and without the physical and prosodic features of face-to-face ignoring, whether an utterance is taken as impolite, as is necessitated by Culpeper's co-constructed view of impoliteness, is hard if not impossible to discern. Bousfield suggests that impoliteness can be categorised as such when it is seen to be:

> "1) Unmitigated, in contexts where mitigation (where mitigation equates with politeness) is required and/or,
> 2) With deliberate aggression, that is, with the face threat exacerbated, 'boosted', or maximised in some way to heighten the face damage inflicted." (p.132)

Finally, this analytic frame was steered by Bousfield's application of impoliteness and power. He suggests that impoliteness can wield power and be "instrumental" when used effectively. He invokes Wartenberg's (1990) three-way distinction of power in this discussion, with Wartenberg's conceptualisation of "influence" (a fine-grained exercise of power where

no explicit verbal or physical threat of violence is used) being particularly pertinent to understanding the power impoliteness had on the recipient, but also on those viewing from the digital side-lines.

Building on the symbolic interactionist ideas, politeness and impoliteness are understood for this research to be linguistic techniques that either maintain, bolster or purposefully damage face. The methods noted above provide a unique lens for understanding the production or subversion of expected interactive etiquette, to better discover what communicative or rhetorical powers they seek to deploy. In building up these discursive methods in this collaborative fashion, a clearer picture can be seen of how hate and counter speech can be understood as genres in their own right, as well as how users manipulate and deploy those genres.

## The Ethics of Digital Data Analysis

The ethical considerations of engaging in social media analysis, particularly fine-grained qualitative analysis, has been an evolving topic of debate as this form of data has proliferated. Social media platforms not only exist as "rich sources of naturally occurring data on any number of topics" (Townsend and Wallace, 2017:190), but because of user generated subcultures and sentinel accounts, as used in this study, those topics are curated, collected and made conveniently available for those interested. As McKee (2013) notes, however, the exponential emergence "risked running ahead of the development of an appropriate ethical framework" (p.299) and the novel medium factors of this new form of data heaped additional extenuating factors onto classical ethical considerations. With social media data being so interconnected, archived and searchable, the ways in which research collect and reproduce data are in need of careful, renewed consideration. Concurrently, however, the associated explosion of criminal, deviant and damaging behaviours (such as hate speech) that have come with this technological data boom necessarily require quick and expansive research to mitigate and ensure protection from harm for those who are newly and more accessibly victimizable.

In their work on the ethics of social media research, Williams et al. (2017, 2017b) and Webb et al. (2016) argue for a more reflexive, more nuanced approach to the moral dealings of online data, placing the rights of the user at the forefront of consideration and moving beyond a purely legalistic ethic. In investigating users understanding and expectation of their rights, with regard to their online data, Williams et al. (2017b) found that the majority expected anonymity in the reproduction of their content and for their consent to be sought out. Both consent and anonymity are at the forefront of the ethical discussion around social media research, and were indeed key factors in designing this research to not only achieve the standards set out by law and ethical committee, but the reflexive, nuanced ethic expounded in Williams et al.'s work.

The data published within this study are done so anonymised and, although attempts at retroactive communication were made, without seeking informed consent. There are no explicit references to the platform they were harvested from, but the content of the messages broadcast remains unchanged. Displaying the message content as found was essential for performing the manner of fine-grained discourse analysis found in this thesis. Particularly since a key epistemological focus was seeking to understand how the different forms of discourse are actually produced and used by social media users, a recreation or bricolage technique would be inadequate. Much discussion is still ongoing about the ethical implications of treating digital data, and social media data in particular, as either public or private. Bishop and Gray (2017) provide a thorough exploration of this idea, discussing the blurriness of defining publicly available platforms owned by private companies as explicitly one or the other, exacerbated by the uncertainty around user consent as it relates to accepting, and more fundamentally understanding, the terms of service they encounter upon sign up. Despite being somewhat dated, the ProjectH research group (cited in Paccagnella, 1997) voted in 1994 on an ethical protocol for not seeking consent for publicly posted Computer Mediated Communications. This protocol insisted that although these data points are collected from privately owned platforms, they are still considered (and understood by users to be) public discourse, "akin to the study of tombstone epitaphs, graffiti, or letters to the editor".

In discussing informed consent, Townsend and Wallace (2017) illustrate the additional logistical difficulty of seeking that consent when dealing with huge numbers of 'participants' who are "unaware of their participation" (p.193). When dealing with logs of archived data this difficulty is amplified because of the possible attrition rates over time due to deleted or banned accounts, especially when investigating a phenomenon like hate speech. While attempts were made to gain opt-out consent from any accounts that were still active and available, it is for this reason specifically that informed consent became impossible to ensure for this study, as using data containing highly policeable speech from years prior resulted in many or most accounts being removed (rendering account holders unidentifiable not only to the researcher but also any reader who may seek to do harm). Additionally, this thesis would argue that any harm that may have occurred in the non-consented reproduction of that content would have been incurred by the sentinel sites from which the data was pulled. At the point of analysis and publication within this research, the labelling of hate speech as such (either racist, homophobic or otherwise) has already been performed by the sentinel sites. The analysis in this thesis does not make any judgements on the content being hateful but takes it as such based on the reproduction by the sentinel.

Regardless of these arguments, this research also took every possible precaution, short of rewriting the messages themselves, to anonymise the data, inspired by the approach taken by Benwell and Stokoe (2006). The reduction of harm is a key component of the ethical consideration of academic research, and while the lengths gone to ensure harm reduction have been discussed, an additional justification for engaging in this area of research is the moral duty to tackling deviant online behaviours such as hate speech, that are in themselves a social and individual harm. Both the pragmatic logistical issues and the theoretical justifications for the ethics which this thesis followed were reviewed and approved by the Cardiff SOCSI ethics committee. In reproducing the data and presenting the analysis in this thesis, every consideration and effort has been made to ensure anonymity, reduce harm and conduct this research in a way that does not restrict the investigation, but stays in line with the standards agreed by the committee.

## SUMMARY

In summation, this chapter has explicated the methodology of this research as a computer mediated discourse analysis (CMDA) of hate and counter speech found on social media. While CMDA as a discipline is expanding and its literature becoming more prevalent, its application to deviant online behaviours, and the response to those behaviours, appears to still be in the early stages of development. The methods of enquiry themselves are split into three distinct but related analytical pathways; using Conversation analysis and Membership Categorisation Analysis to explore the use of online identity, investigating the discourse analytic theories of digital politeness and impoliteness, and taking a constructionist framing linguistic genre and intertextuality, each as techniques and strategies for achieving hate and counter speech as a communicative online action. While each of these approaches varies in its methodological assumptions and its analytic focus, all can be found within the toolkit provided by CMDA, and each are used here as an apparatus to discover ways in which social media users engage in contentious and illocutionarily damaging online discourses.

Crucially, this chapter has attempted to justify both the methodological and ethical choices made in undertaking this research, not simply describe them. In providing a brief narrative account of reaching the decision to define, problematise and investigate hate and counter speech in terms of linguistic genre, this chapter has clarified the logistical and ethical considerations made in producing this research, as well as its data driven motivations. This research cannot claim to be purely inductive in its inception, but the drivers for its design and ethical constraints were led, wherever possible, by qualitative engagement with the data. While this is by no means an exhaustive analysis of the discourses that make up hate and counter speech online, the methodology presented here does intend to provide a novel entry into the groundwork of qualitative online discourse analysis of deviant phenomena, expanding and exercising the toolkit provided by CMDA.

# FINDINGS PART I:

## Hate Speech

Content Warning: This thesis discusses sensitive topics, particularly racism and homophobia. Please be aware that extreme and offensive racial and homophobic slurs will be presented and discussed within the upcoming chapters of this research. For the posterity of the data and its forensic analysis, examples of these hateful rhetorics are presented uncensored and in full as they were originally collected.

- Hate Speech
  - Chapter 4: Genre and Intertextuality
    - Legitimation
    - Excuse Making
    - Projection
  - Chapter 5: Membership Categorisation Analysis and Identity
    - Declarative Identity
    - Oppositional Identity
    - Weaponising Identity
  - Chapter 6: Politeness
    - Markers of Solidarity
    - Going “Off Record”
    - Redressive Action

Fig. 4

# **CHAPTER 4:** GENRE AND INTERTEXTUALITY

## Introduction

This chapter presents an analysis of social media data in which users are making remarks that can be identified as belonging within the linguistic genre of hate speech. The aim of this chapter, having explored the legal and working platform definitions of hate speech and how they combine to form a recognisable genre of speech, is to provide examples of hate speech occurring "naturally" in online settings and illuminate ways in which linguistic techniques are used to manipulate what Briggs and Bauman (1990) call the "intertextual gap" (p.149). Briggs and Bauman suggest that because "the fit between a particular text and its generic model… is never perfect" (p.149) the production of speech or text and its assignment to one genre or another must always necessitate some amount of "intertextual gap". And not only an amount of gap between text and genre necessarily present, but this gap can also be manipulated by the speaker to make the genre of the utterance seem more or less apparent. By extending the intertextual gap between an online statement and the genre of hate speech, the speaker may attempt to reduce its attribution as hateful in the eyes of the audience[7].

While hate speech is widely defined, and policed, online hate speakers are seen to use different speech tactics outside of those definitions to help themselves achieve the communicative aims of the genre. Through a manipulation of the intertextual gap between what is expected or defined as hate speech and what they actually produce, hate speakers are able to maintain a distance from their speech, and achieve their communicative aims in creative and unique ways that may serve to inoculate themselves and their speech from the "racist" or 'homophobic" label.

The three specific linguistic strategies analysed and presented in this chapter are *Legitimation*, *Excuse Making* and *Projection*. Each of these show different ways in which hate speakers are found to play with the conventions of their identified speech genre and modify

[7] Discussed in greater detail in Chapter 2, p.48-52 of this thesis.

how their speech can be read while still achieving the same ends. Briggs and Bauman (1992) explain that in taking a constructionist view of genre "the task becomes one of discovering what portion of the speech economy is generically organised, what portion escapes generic regimentation, and why" (p.139). It is the proposition of this analysis that these three strategies, while being recurrent within hate speech, are deployed to aid the speech in avoiding generic classification as hateful. In using these strategies to reframe, distance or support their hateful rhetoric, hate speakers are shown to do linguistic work beyond the slurs and othering often found in formal definitions of hate speech, and attempt to soften their messaging and broadcast it in a way that is convincing and acceptable to their *intended* audience and to the collapsed probable audience (Marwick and boyd, 2010). With the aims of this chapter outlined, it will now turn to the discussion of the first of the three linguistic strategies: legitimation.

## Legitimation

Legitimation (and authorisation as a specific concept found within legitimation) is discussed at length in the Identity analyses of this thesis (Chapters 5 & 8), viewed particularly through the lens of Membership Categorisation Analysis. However, in this chapter legitimation will be analysed in terms of a linguistic "move" or strategy in the genre of hate speech, or more acutely as a manipulation of the intertextual gap around that genre. In this section, the analysis will dive into how identities are legitimised and give authority to a speaker as a way to distance themselves or to protect themselves from the backlash that may come with producing aggressive and hateful speech.

Identity is very obviously a key component in the use and discussion of hate speech; vulnerable identities being the prime driver to hate speech and a fundamental aspect to its definition and policing. In this section, data will show how the legitimation of identity and of discourse is used to generate an intertextual distance between the speaker and the definitively hateful speech they are using. The key theoretical conceptualisations which will inform understanding of legitimation are taken from Beetham (1991) and van Leeuwen (2007).

Legitimation Example 1

- P1: DO THE PEOPLE OF #Ferguson EVEN KNOW HOW MANY 6 IS?
- P1: LEGITIMATE QUESTION. ASKING QUESTIONS DOESN'T MAKE YOU RACIST
- P2: No, but asking idiotic racist questions sure as heck does.
- P1: I DIDN'T SAY BLACK PEOPLE. YOU IMPLIED THAT.
- P2: No, you implied that in your question. It's a tactic called a "dog whistle". Like innuendo but for racism.

Fig. 5

In the above example, legitimation is very literally spelled out by the hate speaker, attempting to harness that tool to make their language acceptable. P1's initial utterance is a rhetorical response to the general discussion of Michael Brown. The hate speaker specifically refers to Brown being shot 6 times by police, the antecedent trigger event which led to the Ferguson unrest of 2014. As identified by P2, this question serves as a "dog-whistle" (Lopez 2014, Albertson 2014) or a coded message only visible or readable as such by those for whom it is intended. In this instance, the intended audience are those who would seek to de-legitimise the outrage around Michael Brown's death, and by extension de-legitimises the value of those in the community referred to. The initial tweet textually identifies the group they are discussing as "PEOPLE OF #Ferguson", a group who cannot be presumed from the text to be of any specific racial demography. However, the follow up tweet, explicating the assumption that some may read the preceding statement as a racist dog whistle, betrays that the hate speaker (whether intentional or not) understands that it may be read by some as implying that the group whose intelligence he is specifically questioning is the black population of Ferguson. A population who were, presumably, visible to the hate speaker to be outspoken and outraged by the event.

The mocking and attempted de-legitimizing of a hate crime event such as this can be interpreted as being covered under Facebook's first and second tier of hate speech ("Mocking the concept, events or victims of hate crime", "Implies the inferiority of a person's or a group's mental or moral deficiency") as well as Google/YouTube's guidance against denigration of

"victims of a major violent event or their kin". It also serves as a direct reference and justification of deadly violence against black people, which is restated in their response, which falls under Twitter's conduct policy which forbids references to "violent events", "inciting fear" or "content that degrades".

Legitimising this question, even in a blunt and literal way such as this ("LEGITIMATE QUESTION. ASKING QUESTIONS DOESN'T MAKE YOU RACIST"), attempts to remove space for critique and interpretation, instead reframing their previous statement as value neutral and without agenda. Beetham (1991) notes in his discussion of the "Legitimation of Power" that to achieve our communicative purpose one must have "freedom from control, obstruction or subservience to the purposes of others" (p.43). In legitimising their speech and increasing the intertextual gap between what they have said and the genre that the audience member is assigning to that speech, they are attempting to gain freedom from the subservience of the purpose of others. They are attempting to remove themselves from a point of responsibility to abide by the implied or stated social rules (banning racism, hate speech). In the second, clarifying utterance the speaker explicitly assigns legitimacy to their initial question, as well as confirming that legitimacy by denouncing an expected critique. Through this legitimation strategy, P1 is ensuring they are free from "control" or "obstruction" to engage in what they are presenting as value neutral discussion. This utterance manages to achieve the communicative outcomes of the hate speech genre, while placing themselves at an intertextual distance from the specific language tropes associated with that genre. In attempting to legitimise their speech they manipulate the intertextual gap to pro-actively protect themselves from direct, explicit association with the socially stigmatized and policed language genre of hate speech.

Legitimation Example 2

- P1: rabidly racist blacks will not rest until there's a lynching.fuckthem..stand up..
- P2: christ what the fuck is wrong with you
- P3: I blocked him in the past for being a racist
- P1: why are you on my TL? Not happy until you censor me?

Fig. 6

Excerpt 2 shows racist hate speech that falls squarely within any of the above legal or platform definitions. The illocutionary phrase “Stand up” implies and requests a confrontation, invoking a physical opposing response to the chosen target group. Further, the mention of “lynching” provides a specific violent tone, bound up with historical context, against a group based entirely on racial characteristics. This all adds up to a clear example of incitement. The description of “blacks” as “rabid” also fits with the dehumanisation concept central to a majority of hate speech definitions, applying a depiction most commonly associated with dogs. As with the previous example, the identification of the speech as racist by a respondent aligns with the Swalean (1990) understanding of the classification of genre. The speaker is clearly engaging with the generic forms, language and structure of hate speech, even where alterations are made to legitimise their position, and the classification of that genre is co-constructed between the text, the social context and the audience reception. They are successfully achieving the production of generic hate speech because it is identifiable to an audience.

This second example displays an alternative deployment of legitimation, manipulating the intertextual gap so that what can, and is, easily read as racist hate speech can be reframed and defended as free speech and truth telling in opposition of censorship. The hate speaker dehumanizes black people (“rabid”), describes them as acting as a monolith (“racist blacks”), uses a word historically associated with racist killings (“lynching”), and insights anger and physical confrontation (“fuckthem..stand up..”), and yet is able to play with the tropes of the hate speech genre in a way which *can* be read, by those who are susceptible or willing to do so, as legitimate discourse and not racist hate speech.

In their initial utterance the hate speaker presents an authoritative identity of someone standing against an unjust group (invoking Sykes and Matza’s (1957) Condemnation of the condemners), identifying the speaker as a freedom fighter or truth teller, with righteous cause. This suggests that they understand their language to be inflammatory and susceptible to intervention, which is then confirmed three turns later when they explicitly mention the threat of censorship. In invoking censorship, they posture themselves as in one sense legitimate or authoritative (‘what I’m saying is too *real* for some’) and in another as oppressed and an underdog (‘the powers that be don’t want me exercising my right to free speech’).

This legitimation strategy can be identified as van Leeuwen's (2007) second major category of legitimation; "Moral Evaluation" (p.92). What is apparent and identified by P3 as racist language is legitimated through its presentation as a twisted form of moral evaluation. The speaker not only stands themselves as opposed to an imagined racist community (a community who are themselves more likely to be victimised by racists and who are in this very tweet being attacked with hate speech) who are supposedly and hypothetically engaging in violent behaviour, but also positions themselves and their speech as a legitimate victim of unjust censorship.

Sykes and Matza's (1957) techniques of neutralization is also relevant here, in that the speaker legitimizes their hate and their target through the "denial of the victim" (p.668). Here the speaker is re-creating black people, the recipient of their racist, de-humanizing speech, as "rabidly racist" and apparent aggressors in need of retaliation. In doing so, they can suggest that "The injury [they are causing] is not really an injury; rather, it is a form of rightful retaliation or punishment" (p.668).

Legitimation Example 3

- P1: Obama wants more power to act on immigration. HELL NO enforce the fucking laws on the books asshole. I am fed up that nigger in the WH
- P2: I'm embarrassed this racist is from the state I work in…
- P3: you had me there. Right up to the bit you were racist.

Fig. 7

This third example shows P1 legitimating their speech by attempting to extend the intertextual gap between racist text and hate speech genre. By opening their utterance in a way that frames their speech as political debate and critique of leadership, the use of an outright slur at the end of the message is justified and legitimised because it is seen as being driven by the stated critique and not because of racial hatred.

Both respondents are able to easily and rightfully identify the hateful speech as such, but P3 specifically explains how the message was received by opposers of racist language as being

an attempt at reasonable debate that is retroactively tainted by the inclusion of explicit racism. Were this message received by someone whose political or social values were less opposed to racist slurs, its use may well have been successfully legitimised by being couched within a form of justifiable debate and critique. While the opening two sentences may well have been intended as fair critique of a politician, the use of the word “nigger” strongly suggests that the assessment is informed at least in part by racial stereotyping. Additionally, the specific complaint that they are “fed up of that nigger in the WH” (White House) brings implicit connotations that people of that kind, those who would be described by this speaker as “niggers” do not belong in the highest political office of the United States. These implicit dog-whistles are no doubt understood as targeted racially motivated attacks, the respondents shown both attest to as much. However, were this message received by someone who does not find the statements as morally reprehensible, the connection between a desire for more power, an inaction on law enforcement (and by extension a tacit endorsement of lawlessness) and black people in the White House are both visible and legitimised. This legitimisation process not only serves to make the use of racist hate speech more palatable to those who may be in the audience, but to also distinguish the speaker as at a moral distance from the hate speech they are producing. The manipulation of the intertextual gap here is deployed to show that the hate speaker is not engaging in the language, structure and communicative goals of hate speech because they are themselves a hateful racist (even though that is what is demonstrably seen by the respondents) but because they are engaging in legitimate political discourse.

Legitimation Example 4

- P1: If you're a Christian or if you care where this country is going. Boycott Burger King. #RainbowWhopper #Adam&Eve NOT #Adam&Steve
- P2: [P1] is tweeting anti-gay white trash crap – please troll this closet queer who loves his "god" so much
- P3: LMAO at these insecure FAGGOTS who can't accept that WE are entitled to our own opinions #sueme
- P4: Jesus will love you more for not eating a salty meat patty because… wait, because why?
- P1: When did I say anything about Jesus loving me more? It's not about the burger. It's about what I believe.

Fig. 8

Excerpt 4 shows the legitimation of speech, and an invocation of authority, through the martialling of both religious status and nationality. P1's homophobic hate speech, employing mocking hashtags and denouncing a company engaging in gay pride promotion, is legitimated through the idea that authority to expound such hate is granted by association with a religion or a country. The initial utterance contains no explicit slurs but portrays a clear bias against a group based solely on their sexuality. The specific invocation of the Christian religion, a religion that has a dogmatic history of condemnation against gay people, and of a non-specific "country" that presumably has a large or dominant Christian population, stands the speaker (and anyone who may share this religious or national identity) as a part of the standard identity category. The conflation of religious morality and national identity, particularly when bolstered by hashtags referencing the "rainbow" (a symbol of the gay pride movement) and the anti-gay adage "Adam and Eve, not Adam and Steve" (a reference to the biblical creation story often used to demonstrate God's preference of heterosexual relationships and his condemnation of homosexual ones) paints a clear picture in which gay people, and corporate support for gay people, are understood to have a negative impact on "where this country is going" and be an attack against the expected standard of Christian morality. All this is achieved without the use of slurs or explicit personally motivated attacks against gay people.

In deploying these forms of legitimacy and manipulating the intertextual gap to achieve hate speech aims without the use of the expected hate speech tropes, P1 is able to broadcast their messaging effectively to those who would hear it. This is confirmed in both denunciation from P2, who brands P1's speech as "anti-gay white trash crap" and in amplification and explication by P3 who deploys the homophobic slur "Faggots" while proclaiming and defending their shared right to be entitled to an opinion.

When the logic of P1's messaging is questioned by P4, who inquires as to why the choice to frequent or boycott a burger restaurant would have any meaningful connection to the person's favour with God, the hate speaker rebuffs their suggestion and reframes their initial message as one of "belief". Again, the hate speaker legitimises through the use of religious language and works to remove or reduce their responsibility to their messaging. In stating it as a "belief" which aligns with their Christian values, the hate speaker implies that those beliefs are both religiously sanctified and protected from criticism. Not only is the speech supposedly protected from criticism because of their apparent foundation in religious doctrine, it also reframes the understanding of the hate speech not as an outward act of aggression toward a vulnerable group, but an inward belief that does no harm (Similar to Sykes and Matza's (1957) Denial of Injury). Despite the initial utterance working to convince hearers against gay people and gay acceptance, the hate speaker reframes their speech as a purely internal belief that is of no consequence or business of others. In each of these different dimensions, the hate speaker is manipulating the intertextual gap and removing their personal responsibility further and further from their speech. While they are demonstrably seen by both collaborators and opposers to be achieving the communicative aims anti-gay hate speech, their rhetorical strategies protect the speaker and legitimise their messaging.

Legitimation Example 5

- P1: I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey
- P2: I can proudly say that as a human being that I hate pathetic racists like you
- P3: You have to give children a little leeway when they say dumb, racist shit.

Fig. 9

This final example of legitimation involves very explicit hate speech, complete with the deployment of two anti-black racial slurs, and the use of violent profanity to attack and dehumanize their chosen target, the at the time president of the United States, specifically for their status as a member of a marginalised group. P1 provides no outside analysis of what they find upsetting about the president, merely that they "hate" him and classify him as a "nigger" and a "stupid porch monkey".

The speech used here is very clear and apparent hate speech, which is identified and confirmed as such by both respondents, however P1 still insists on an attempt at legitimising that speech through their identity as an "american" acting "proudly". This attempt at legitimisation is identified and repackaged by P2, who expands on their legitimation strategy to provide themselves with a greater sense of authority in their retaliation. While the legitimation strategy used by P1 was not strong enough to fully achieve that rhetorical practice, at least not with these two respondents, it can be read and understood what the intention of that strategy was, to manipulate the intertextual gap and distance the hateful speaker from the hateful speech. This torrent of slurs and profanity is not being generated by a hateful racist, but by a proud American, and this kind of speech would be used by anyone else who identified as such. By invoking a specific group alignment, P1 can be understood as using Sykes and Matza's "appeal to higher loyalties" (1957: 669) as a means of neutralizing their guilt or responsibility and legitimizing their speech. The generalized social norm to avoid explicit racism is sacrificed in favour of their perceived responsibility to their "American" group membership. P1 is achieving not only the communicative aim of hate speech which seeks to attack and dehumanize a member of a vulnerable group, but also to legitimise those ideas with the hope of propagation and broadcast. This is done by manipulating the intertextual gap to draw on genre tropes of nationalism and patriotism as a means to

authorise their explicit hate speech and reframe it as something acceptable, something expected or even something required.

Each attempt at legitimation, whether successful or not, manipulates the intertextual gap between the communicative aims of the hate speech genre, the actual text of what they are producing and the person producing it. Whether explicit or implied, each example above is read as belonging to the hate speech genre because of its intertextual connection to the rhetorical tropes and communicative aims of that genre. The manipulation of the intertextual gap around that genre, through the choice of legitimising rhetoric allows hate speakers to work visibly within the genre of hate speech while still maintaining (or attempting to maintain) some level of protective distance from their speech. By legitimizing their speech through "moral evaluation", "rationalization", or "authorization" (van Leeuwen 2007:92) the hate speakers are able to work within the intertextual gap between the definitional tropes of the genre and the communicative aims they seek to achieve. Legitimation in this way provides a protective shield that hate speakers can use in an attempt to keep labels like "hate speaker", "racist" or "homophobe" at bay while producing and broadcasting a genre of speech which may attract them.

### Excuse Making

Excuse making is another form of justification that appears frequently in the hate speech data collected for this research. Tedeschi and Reiss (1981) suggest that:

> "Excuses are explanations in which one admits that the disruptive act is bad wrong or inappropriate but disassociates himself from it. Justifications are explanations in which the actor takes responsibility for the action but denies that it has the negative quality that others might attribute to it" (p.281).

Here their definition of "Justification" follows closer to what in this chapter is referred to as "legitimation", and justification is used as an umbrella term to account for linguistic techniques that explain for an action's (verbal or physical) happening through legitimation or excusing. However, the definition of excuse making hints towards the distancing of the

intertextual gap between speaker (or speech) and genre, creating a situation in which the speaker is justified in their use of admittedly incorrect speech because they have disassociated themselves from it.

In their work on excuse theory, Mehlman and Snyder (1985) identify three discrete operations which allow the excuse makers to avoid the judgement of engaging in "poor performance" (p.994). They suggest that:

> "Generally, excuse making reflects efforts to raise consensus and distinctiveness information and to lower consistency information that is relevant to an ego-threatening event" (p.994)

Firstly, consensus raising refers to the practice of suggesting that anyone experiencing the same situation would perform equally as poorly, thereby reducing personal or individual responsibility. Secondly, distinctiveness raising describes the process of making the individual situation that this instance of poor performance occurred in distinct from others, and therefore not generalisable. This works to separate the person performing poorly, from the person in all other situations in which they are presumed to be virtuous. By enforcing this separation, any judgement that is given for a poor performance is seen to be directed specifically at the performance and not at the person. Lastly, consistency lowering refers to indicating that one's level of performance wavers across activities and times. This suggests that a good or poor performance is due to external factors and not a reflection of the individual, compared against someone who acts consistently well or poorly across many instances.

In each of the examples of excuse making analysed in this chapter, the "performance" that will be discussed as having been done poorly is the production of socially acceptable speech. The misstep or poor performance that speakers are being seen to undertake is broadcasting speech outside of the expected norms of the community, namely racist or homophobic hate speech. When hate speakers perform their genre of language, the presence of excuse-making signifies their understanding that they are performing social norms poorly

and that they must augment their speech to mitigate that performance, manipulating the intertextual gap between their selves, their speech and their chosen genre.

Excuse Making Example 1

- P1: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.
- P2: hahahahahahahahahaha does that apply to Disney movies? Or making S'mores? What about birthday parties??
- P3: white people have that effect too

Fig. 10

This first example employs what Briggs and Bauman (1992) might refer to as a "generic framing device" (p.147) in the use of the phrase "I'm not racist but…". Regardless of the intention of the speaker, this phrase will "unleash a set of expectations regarding narrative form and content" (p.147) because of the known associations of this phrasing with the genre it denotes. This harkens back to the primary notions of intertextuality, in which specific tropes or turns associated with a genre are understood through people's previous engagement with those texts and discourses, and in choosing to embrace or negotiate those genre tropes one is able to manipulate the intertextual gap between their produced speech and the genre of speech they are invoking.

The myriad social media profiles (@YesYoureRacist on Twitter, r/im_not_racist_but on Reddit) and internet memes dedicated to the idea that this phrase is a pre-cursor to upcoming racism demonstrates its acceptance as generic nomenclature. However, following this phrase two examples of consensus raising are performed, attempting to remove the individual responsibility for their racist assertion and disperse it all who are reading. "…[E]ven you cant deny it" and "Don't even try to deny it" shift the level of responsibility on the speaker and manipulate the intertextual gap between what is visible as a racially motivated judgement and the genre it falls into. By asserting this as an agreed, universal notion based on consensus, P1 has tried to distance this utterance from the genre of hate speech and reframe their individual poor performance as a common issue. Their use of consensus raising reorients the

speaker as working within a controversial, but more importantly an acceptable or necessary genre of speech. In this way they are portrayed as a 'truth teller' describing the universal experience that others are too afraid to say themselves.

P1's utterance here also provides a clear example of Sykes and Matza's (1957:667) "denial of responsibility", providing a spurious excuse for why they are making the racist statement they are making. They are not proffering this suggestion because of their own racist tendencies, but because of forces outside of their control impacting their judgement. That force being the supposedly inarguable fact that "black people make everything scarier". This statement manages to be both a racist assertion, and a neutralising excuse for itself simultaneously.

Excuse making and explaining are viewed here as tools to manipulate the intertextual gap through the use of hedging devices. Where text is known by the author to be identifiable as socially unacceptable, as is the case with hate speech and prejudiced language, pre-emptive excuses and hedging language are deployed to distance oneself from the responsibility of broadcasting that language. In manipulating the intertextual gap and reframing speech that would be generically defined as hateful into something that may arguably belong to another, more acceptable genre, the speaker can reasonably claim (or hopes to claim) that they are not creating hate speech.

Excuse Making Example 2

- P1: I don't like to see cops beat honest men for no reason, but I do like to see them shoot nigger criminals for good reasons. #Ferguson
- P2: <---- attention whore
- P3: Racist cunt, asshole, attention whore, and flaccid.

Fig. 11

This second example, despite engaging in explicit racist language, attempts to excuse its hateful content through raising distinctiveness between two theoretically separate situations. What is produced here shows a very explicit example of Mehlman and Snyder's

(1985) “distinctiveness raising”, in which the hate speaker offers a definitive example of where his judgement can be seen as virtuous and acceptable. The initial situation in which P1 asserts he acts correctly (“performs well”) is used to illustrate a contrasting distinction from situations where he acts in discordance with social norms, as in the second situation. The implication then is that one cannot judge the speaker *themselves* as being bad based on this utterance, but that in this specific case they have acted badly. Their socially objectionable opinion is presented as an outlier and not indicative of the totality of his analysis of police misconduct. They are not making this opinion based on racial stereotypes, or a desire to see police violence, because their initial utterance explains that where a man is unjustly beaten, he does not enjoy it. This individual, second utterance is a distinct one, and justified as being acceptable to espouse by the first.

This example also illustrates the consistency lowering strategy. By demonstrating an inconsistency of reaction in different situations this suggests that the speakers apparent lack of care is not an innate personal trait but is determined by extraneous factors. In doing so, the intertextual gap between the produced speech and the hate speech genre is manipulated to give the speaker deniability from having invoked it. Through this strategy the speaker is asserting that their unacceptable talk is not motivated by racial hatred, but by an acceptable analysis and discussion of laws and consequences. This ‘analysis’ they provide is bolstered by their “denial of the victim” (Sykes and Matza, 1957), when they remake the victim not as an “honest man” but as a “nigger criminal”. They are not only excusing the use of an unacceptable genre, but attempting to reframe their speech as a different genre entirely; the use of the slur is not an indicator of the genre of speech he is using but an irrelevant language choice made in his pursuit of a reasoned discussion on police conduct.

Excuse Making Example 3

- P1: “Gosh, those fat black girls sitting over at that table sure are quiet.” – No one in the history of mankind #fatblackgirls
- P2: So, brilliant one: is it their ethnicity or their weight that causes them to be so loud? Show your evidence.
- P1: don’t let your white guilt get in the way and also I bet I have more black friends than you
- P3: “I bet my relative proximity to black people in the area somehow proves I’m not racist!”

Fig. 12

The third example shows once again a take on consensus raising. In this example P1 uses sarcasm and irony in designing a mocking assertion and framing it in a “meme” format. A recurring joke on social media platforms is to invent a quote and attribute that quote to a person, group or as is the case in this example, to no one, to confirm or explicitly reject what is stated in the quote. This particular example uses the meme template to consensus raise by sarcastically producing an apparently unbelievable quote (that “fat black girls” are often quiet) and attribute it to “no one in the history of mankind”. In suggesting that “no one” would ever be motivated to make such a statement, the implicit messaging is that everyone would agree with the opposite and so raises the consensus of that message. The consensus that the hate speaker has raised here is the assertion that everyone would agree with the sentiment that “fat black girls” sat at a table are loud.

The quote and its hypothetical rejection make sweeping generalisations about the behaviour of “fat black girls”, and this racist, misogynist, fat shaming hate speech is identified as such by P2, who attempts to enquire which of the featured characteristics the hate speaker suggests is responsible for their unacceptable loudness. This questioning motivates a response which works to excuse the racist assertion through raising distinctiveness of this instance of poor performance against other examples where P1 is demonstrably *not* racist. In making the statement that they “have more black friends than you”, the hate speaker is setting the current instance where their rhetoric has been identified as racist as a distinct and

individual occurrence, that must not be indicative of their person or their routine behaviour because of their exemplary relationships with black people. They must not be racist because they, supposedly, have black friends who do not object to their statements. They are also performing Sykes and Matza's "condemnation of the condemners" (1957: 668), rejecting the criticism and implying hypocrisy and insincerity by claiming their intervention is driven by white guilt rather than altruism. Here, these examples of excuse making manipulates the intertextual gap in a way which does not distance the hate speaker, or remove their responsibility for their language, but justifies their decision, or their authorisation, to engage in this language. Sweeping generalisations about a race of people are not prejudice or hateful when produced by this person, because the broader population all agree with it (they suppose) and if they don't, this instance must be distinct and non-indicative of their general behaviour, because of their black friends.

Excuse Making Example 4

- P1: I don't like public display of homosexuality , it offends me and the Majority! Keep it PRIVATE!
- P1: Isn't your "husband" with you? Have him take you2the Zoo, you will both learn something, No Gay Animals
- P2: he's so cute when he's being ridiculous.
- P1: There you go, again! I have nothing against Gays, as long as you keep it PRIVATE! Marriage is PUBLIC!

Fig.13

In example 4, P1 makes an explicit invocation of consensus raising, directly aligning their views with an assumed majority opinion. The hate speaker denounces "public displays of homosexuality" as offensive, demonising a vulnerable minority group. In their statement they presume and assert their values to be those held by the majority and in professing their majority in-group status use that to justify and excuse their proclamation. However, inversely, the pre-emptive use of this consensus raising justification suggest an understanding that their statement is, if not offensive to the presumed majority, liable to receive backlash from a vocal minority, one whose offense and personal values require a majority consensus to effectively

argue against. In invoking this presumed majority, the speaker works to distribute responsibility for their proclaimed beliefs among a great number of people, an asserted number of people who must be larger than the number who may find it offensive.

The excusing work is bolstered once again in response to sarcastic counter speech provided by P2, when the hate speaker announces that they "have nothing against Gays", using consistency lowering to suggest that this reprimanding of "public" gayness is not indicative of a consistent hatred towards gay people, but is instead rooted in the innate distinction between marriage(presumably heterosexual marriage) as a public event and homosexuality as a private event. Their admonishment of gay people expressing their love publicly is not hate speech because it is not driven by a dislike for gay people, they suggest, it is because there is a majority confirmed difference between the arenas in which straight marriage and gay love may occur.

Interestingly P1 also deploys a form of van Leeuwen's (2007) legitimation through authorization, particularly "theoretical rationalization" (p.103), exemplifying an imagined authority on sexuality from the animal kingdom. Van Leeuwen's theoretical rationalization suggests that people legitimize their assertions by discussing them in reference to them being a "natural" truth or "the way things are". In stating that there are "No Gay Animals" the speaker seems to imbue non-human animals with some form of naturalised expert authority on the subject of sexuality and presents that as a legitimation for their assertions. Whether this statement is demonstrably true or not, it references an assumed and expected natural order of "the way things are". This works in tandem with the excuse making to manipulate the intertextual gap, moving themselves and their speech further from the hate speech genre, while working within it in a way identifiable by other interactants. Their speech is read by others to be hateful because it does deal in some of the features of generic hate speech (making derogatory and demeaning statements about gay people), but is also couched within rhetoric that attempts to paint their speech as declaration of majority agreed values and a discussion of shared cultural norms, genres of speech which are not objectionable and worthy of policing.

Excuse Making Example 5

- P1: I support gay marriage but 2 nigga's kissing in public should be illegal keep it for when your alone fuck ! Its disturbing !
- P2: What about two straight people kissing in public?
- P3: ok then, lets make It illegal to kiss your girlfriend in public as well, sound good?
- P4: Your lack of understanding about apostrophes is disgusting.

Fig. 14

This final example shows P1 engaging in consistency lowering as an excuse to distance themselves from association with the hate speech genre they are accused of working within. The hate speaker makes the assertion that two men kissing in public should be made illegal and that their witnessing of the act is disturbing. However, P1 attempts to mitigate this clear and obvious demonisation of the right of gay people to express their love publicly by contrasting that with the proclamation of their apparent support of gay marriage. In demonstrating the inconsistency between their disgust at gay people expressing themselves publicly and their support of their right to marriage, the hate speaker shows that their act of poor performance is not indicative of their actions at all times, and thus is not a reflection of the homophobe they are accused of being. They cannot be engaging in the linguistic genre of hate speech because they are outwardly professing their support for gay rights, this apparently and presumably homophobic opinion must be based in some other motivation.

According to Tedeschi and Reiss' (1981) definition of excuses, the unprompted production of an excusing strategy such as this is an admission by the speaker that they are aware that what they are saying is bad or will be taken as poor performance and is in need of a caveat to justify that performance. In asserting their support for gay marriage, an institution that is both legally accepted and normatively accepted by the majority of their assumed audience, they manipulate the intertextual gap between themselves and the other offending part of their speech. They are not performing hate speech and are not themselves hateful, because of the excuse they have provided. Their statement demeaning gay people for their

public expression of love must be driven by something else, namely a discussion of acceptable public versus private behaviour.

In each of these examples, excuse-making is used to manipulate the intertextual gap to disassociate the speaker from the genre of speech they are engaging in. In some cases, they dilute their responsibility by framing their opinion as being held by an assumed consensus, and in others they reduce their personal connection with their speech by displaying inconsistencies in the history own their own behaviour or character, reformulating their hate speech as a situational occurrence rather than an individual character trait. In each instance though, they attempt to control intertextual gap between themselves and their text, leaving room within that gap for interpretation and rationalisation of their speech.

## Projection

Projection is another linguistic technique through which hate speakers are able to manipulate the intertextual gap between the content of their utterances and the hate speech genre they work within. What will be described and explained here as "projection" removes responsibility for the hateful content that is invoked by the hate speaker by placing responsibility or agency for that content onto their perceived audience through questioning their audience, or phrasing their speech as to suggest the audience created it or has agreed to it. Through a few different linguistic techniques, hate speakers are able to turn their own individual hateful views into co-constructed collaboration with those presumed to be reading their output. This final strategy deals mainly in two specific linguistic theories, Directive Speech and Idealized Cognitive Models. To properly frame the upcoming examples, a brief discussion of those concepts will be made now.

### *The Power in Projection*

Hernández and Mendoza (2002) discuss the illocutionary force of directive speech and the ways in which language is used to cause action in or by another person. Hernández and Mendoza begin with an acknowledgement of the concept of "illocutionary scenarios" as proposed by Panther and Thornburg (1998), which function as a "type of generic knowledge

organisation structure" (p.260). When an illocutionary scenario is created it presupposes the felicity conditions, the capacity and the willingness of the addressee to engage with and complete that illocutionary scenario. Their example "will you close the door" presumes the presence of a person, a door, the ability to close the door, and the request or directive speech itself invokes the generic structures and features of the illocution it is invoking. In an example more pertinent to this analysis, illocutionary scenarios requesting an addressee to give permission for, or make active acceptance of, things like terrorism or mass immigration the generic structure of this illocution and the hate speech genre are invoked and presumed in its use. Additionally, the felicity conditions of a listening recipient, a definite threat from immigrants/terrorists, the ability to be an active participant in the acceptance or refusal of their impact, are all presupposed at the point which an illocutionary scenario, or a projecting directive speech is made.

Hernández and Mendoza also discuss the role of optionality in requests or questions, as opposed to orders or proclamations. In phrasing directive speech as a request that the addressee is free to accept or deny, they suggest that "the speaker minimises his importance at the same time that he maximises that of the addressee, thus increasing the degree of positive politeness of the act". This increased positive politeness decreases the likelihood of refusal to the request and this shift in importance from the speaker to the addressee suggests a manipulation of the intertextual gap. The focus in this kind of directive speech places the responsibility for the rhetoric within the utterance upon respondent and not upon the person asking the question. In this sense, a controversial question is only a question, and it is the response to that question which is the moral focus. Hate speakers can then work within the genre of hate speech but distance themselves by placing the agency of the moral upon their presumed respondent.

Hernández and Mendoza (2002) expand the politeness principle with the idea of the "Idealized cognitive model" (ICM) of cost-benefit (p.268). The concept of ICM is taken from Lakoff's (1987) work and describes a gestalt structure of knowledge, taking what information is presented and building a coherent cognitive model from it. Addressing some issues with classicist criminology and rational choice, this model accounts for a lack of information and understanding, suggesting that an idealized model of a situation is created based on what

information is available, and this is used to develop a cost-benefit analysis by which a decision is made. Hernández and Mendoza suggest that from an ICM designed by the information presented, politeness may compel an addressee to perform an action which will bring benefit to the speaker (or other) when they are able. This does not account for addressees or members of the assumed audience who have additional information on which to create the ICM, but when those uninitiated are presented with presumably correct information of a distressing scene (e.g., terrorism, immigration), a 'knee-jerk' reaction may be elicited, based on this as the totality of their ICM of the situation.

Woofit (2005), informed by Billig (1990) discusses the implicit argumentative nature of discourse, suggesting that persuasive rhetoric is always present in talk. The projection of agency in the co-creation of hate speech upon an assumed reader is not done without rhetorical influence. The talk provides resources for how to understand the world and the topic being discussed, and those inform and persuade regardless of the choice or optionality being presented. When someone is presented with a yes or no question about a complex topic (like terrorism, immigration, race relations etc), Hernández and Mendoza (2002) suggest that it "restricts the potential subsequent conversational moves to two: refusal to comply with the request and agreement to comply with the request" (p.276). The projection of 'choice' reduces a complex conversation to a binary, which along with the influence of the ICM of the information given and the persuasive nature of talk, skews the impartiality of that choice.

Hernández and Mendoza also discuss the role of threat in directive speech, noting that threats are often deployed when a resistance to a request (or "blockage" p.277) is anticipated. Indirect directive speech (questioning) as discussed above is often used in place of ordering when a hierarchy of power is not present, or not readily identified. Threats can be used to explicitly state the cost of non-compliance that would be implicit when receiving an order from a superior. In increasingly anonymous social media interactions, that power structure is often vague, and so an explicit statement of cost increases persuasion. The implicit cost of accepting a racist assertion is the possible public shame of being identified as a racist, but the explicit cost presented by the indirect threat (terrorism, cultural erasure etc) may appear greater to the respondent. The cost presented by the threat may even more

persuasive because it is presented as an unavoidable alternative, when compared to the implicit cost.

Additionally, by framing the explicit threat as one to the community, not only the individual, the speaker reorients the threat as *socially* beneficial rather than individually. This is not a self-centred threat, but a socially benevolent one. Potter's (1996) notion of "Stake Inoculation" suggests that this can protect one's arguments from being dismissed as being driven by egotism. The speaker's individual safety is not at stake, and so there can be no perceivable bias behind their threat. This effect can also be produced by presenting oneself as informed, rational or personally distanced from a topic.

The following examples will all show instances of "projecting" speech, in which a hate speaker manipulates the intertextual gap between their speech and the genre they invoke by implementing directive speech as a means to shift agency to their perceived audience.

### *Projecting Data*

Projecting Example 1

- P1: Im not racist but if you think that everything should be handed to you based on your ethnicity then get the fuck out of the land of the free
- P2: Like if you're white?
- P3: You just described every straight, white man that I've ever met.

Fig. 15

Example 1 shows the projection of a racist utterance, invoking stereotypes of entitlement and exceptionalism and placing the agency for confirming those stereotypes upon their perceived audience. After excusing their forthcoming utterance by announcing they are "not racist", P1 creates a caricature of an 'other' who believes they are entitled to everything because of their ethnicity. This othering also performs a "denial of the victim" (Sykes and Matza, 1957), characterizing them as entitled and worthy of retaliation instead of

the victim of racist stereotyping. Because no other information is presented this may be the totality of the ICM upon which the reader should make their agreement or rejection. The reader may be inclined to believe that there are in fact people who think "everything should be handed to you based on your ethnicity" because of the information they have been presented and the positive politeness with which they are influenced to produce.

This comment is not framed explicitly as a question, but as an ultimatum statement which requires acceptance or refusal by those who would read it. This ultimatum framing effectively reduces the discussion of race, entitlement, equality and equity down to a binary choice. This is reinforced further with the implied threat that standing counter to P1's suggestion should result in your expulsion from "the land of the free". This is a particularly noteworthy synonym used for the U.S.A. because it invokes assumptions of American exceptionalism, but also the intrinsic value of "freedom" purportedly exemplified by the U.S. Presented in contrast to the idea that there are entitled people who believe they are deserving of special treatment based on their race, this reinforces the incongruence between these patriotic ideals, and the supposition that those who think that way do not belong. This bolsters the rhetorical force of the statement, to convince the presumed reader to adhere to ideals that are presented as in keeping with "the land of the free".

In projecting this statement, using the phrasing "if *you* think", P1 is manipulating the intertextual gap between themselves and the use of the hate speech genre. Even though the speaker identifies in the outset that their messaging may be read as racist, they are presenting this as a request for the reader to confirm that entitlement based on racial identity is not in keeping with the ideals of their country, and not an attack on the caricatured stereotypes of anyone who would not fit the white American archetype that the speaker expects to uphold the values of the "land of the free".

Projecting Example 2 & 3

- P1: This video tells you everything you need to know about the state of London and the UK in general. [VIDEO]
- P2: Is it even worth calling it the U.K. anymore? It's no mans land.
- P3: They want to delete your heritage and culture and you are supposed to be polite about objecting to it.

Fig. 16

The two examples analysed here are uttered by P2 and P3 in response to P1 sharing a YouTube video link titled "Sadiq's Stooges Try To Eject London Assembly Member for Saying Labour Isn't Patriotic", in which the London mayor and his party are described as "unpatriotic" and "not liking" Britain. That first utterance is shared by a high-profile right-wing commentator, while the responses are anonymised accounts.

The first response made by P2 uses the context provided by the initial utterance and the video material provided to design a question for his presumed audience that draws into question the identity of the country that is being discussed. The suggestion that a country is no longer what it was, or worth being referred to by its name because of modern changes are a frequent "dog whistle" tactic to perform covert racism (Albertson 2014, Lopez 2014). Rather than merely stating the assertion that 'it is no longer worth calling it the U.K. It's no mans land', P2 phrases their notion as a question in an effort to direct the audience to engage and answer it. As noted earlier, with anonymity reducing, or obfuscating the hierarchical power to simply make a statement and have it accepted, the speaker here frames the idea as a question to be confirmed, alleviating themselves of the responsibility for their speech and leaving it with whoever responds. It is the active agent who engages and confirms this question who bares the responsibility and the morality of the assertion.

By moving the emphasis away from the speaker and finishing the utterance with the threat that the U.K. is 'no mans land' (a stateless, lawless land), P2 circumvents the issue of lack of power over the listener, adding more persuasive, illocutionary weight to their question. This dog whistle racism against the supposed non-U.K. 'other' turning the country

into “no mans land” is projected onto the reader, making them the active agent to confirm or deny this question. This manipulates the intertextual gap between the P2 and the hate speech genre they are working in by asking questions of nationality and statehood, rather than espousing racist, nationalist rhetoric. They are not bigotedly stating that the U.K. has been deformed and changed by foreign influence and non-British others, but are asking an assumed reader if they believe their country is still recognisable as “the U.K.” in its current state.

P3 takes a similar tactic, but instead of framing their utterance as a question, they present their argument as a protective explanation of the threat imposed on “you” (the audience presumed to be engaging). In doing so they implicitly remove themselves as being in danger from this supposed attack, this announcement is purely for the protection of the “you” proposed to make up the majority of the U.K. In using the pronouns “They” and “You” the hate speaker projects their statements and insists that a personal destruction of *your* culture and heritage is at stake from “them”. By twisting the threat away from themselves and onto an assumed reader, they perform ‘stake inoculation’ to protect themselves from accusations of bias or hatred. The statement is being made for the protection of the vulnerable “you” that they are speaking to.

Again, the reference to culture and heritage is a common dog whistle tactic, used to alert supposedly legitimate members of one race or nationality that they are under threat from a group whose culture and heritage must be at odds with and destructive towards your own. Instituting these dog whistles and vague references to “you”, “they” and “the U.K.” provides the speaker with a convenient safety net that no explicit mention of race was made, again distancing themselves from the hate speech genre they are invoking. The “they” invoked is not a particular race, religion or nationality, but a convenient identity category containing those who would do “you” harm. The split is not between nations and races, but between heritage and those who wish to destroy it, the speaker suggests. This can be seen as another form of “blockage removal”, eliminating the barrier for readers who are consequently more likely to (or at least less restricted from) agreeing with statements that demonise and cause illocutionary harm to a vulnerable group.

By projecting the destruction of heritage and culture as a threat to an assumed "you", rather than the speaker themselves, the intertextual gap between the speech and the hate speech genre is manipulated allowing room for interpretation as an act of protection from violent aggressive forces, not a demonisation of a foreign "other".

Projecting Example 4

- P1: #blm How comes none of you protest Jay Z and Cardi Bs use of the N word? Why don't you address the anti Semitic views in the black community?

Fig. 17

Example 4 shows a speaker once again projecting their speech onto an assumed audience. Here P1 is attempting to undermine the Black Lives Matter (BLM) movement by using projecting questions to highlight apparent shortcomings of the black community. In doing so, P1 works to demonise the black community for its use of "the N word", its suggested anti-Semitism, the hypocrisy for treating them both as unencumbered entitlements of the black community, and invalidate the primary concern of the BLM movement, the unjust and over-representative killing of black people by the U.S. police.

By beginning their message with the hashtag "#blm" they make themselves available to a wide audience of not only supporters of the movement, but also those who may be using the hashtag for critical discussion as is seen here. In using this hashtag, P1 may assume a prospective audience made up of people arguing from both sides of this ideology exposing themselves to those they hope to convince, those hope to undermine and those who may confirm and back up their sentiments. In this example, P1 makes the projecting accusation that there exists a double standard between the condemnation of the use of racial slurs and anti-Semitism in the "black community" and those who are not.

The first sentence of this utterance asks why Jay Z and Cardi B, two high profile black rappers are allowed to use the N word without reproach from the presumed audience. The assumption within this sentence is that the reader, or those who support the BLM movement,

would protest the use of the N word by those who are not categorically similar to these musicians, i.e. not black. This, the speaker implies is an inconsistency that is demonstrable of the privilege afforded to black people.

P1 makes no explicit reference to race when making this initial question, so the decision to frame this discussion around two famous musicians could be stretched to be interpreted as a discussion of status, wealth or profession as rappers. However, its placement alongside the BLM hashtag and the invocation of "the black community" suggest this is specifically a discussion around the consent for black people to use the N word freely, where members of other races would be lambasted. With this initial inconsistency alluded to, P1 moves on to question why anti-Semitism within the black community, a judgement they assert is held by "the black community" as a monolith, is not addressed. The implication here is that anti-Semitism is agreed upon as a social ill, but that its 'badness' is only addressed when performed by members of other races. While the demographic character of the speaker cannot be assumed with certainty, the positioning of "the black community" as containing these inherent failures and as in need of discussion suggests that the speaker believes that community are afforded privileges that other races are not, particularly by those who support the Black Lives Matter movement. This implies the supporters of those movements are invalid, or at least hypocritical and inconsiderate of the apparent innate failures of that community.

In framing these ideas as questions rather than statements, the speaker projects their beliefs upon the presumed audience, and displaces the responsibility for the morality of their content. The speaker is not critiquing or condemning the black community and its apparent inherent failings, but rather dispassionately presenting one side of a debate that the reader must confirm or deny. It is the responsibility of the reader to explain if they believe these injustices are acceptable to ignore, or if not become the active agent in validating them. Based on the information presented by P1, use of the N word and anti-Semitism are taken for granted as being morally repugnant, used by the black community and occurring free from condemnation. With no additional information this ICM can be taken as the totality from which the audience must evaluate and make their decision. From this ICM, a reader could assume that anti-Semitism is bad and it is unquestionably used by the black community, and

with this knowledge it would be very difficult to rationally disagree with the suggestion that these issues should be addressed. The persuasive rhetoric in this utterance is weighted strongly to make the choice seem obvious and morally correct, instead of a sweeping generalisation, void of context, about black people as a monolith. The inherent threat of this question is to be viewed as someone who overlooks anti-Semitism and is hypocritically preferential to "the black community" if you refuse the call to investigate. This threat reduces the resistance, or removes the blockage, to refuse or question what is being presented in the ICM.

Projecting Example 5

- P1: It's typical of the #BBC to label people who don't want Wales to be turned into a refugee camp 'extremist'. How many economic migrants do we need to accept before they are satisfied? #DefundTheBBC
- P2: [UK FLAG][WALES FLAG]Wales relies heavily on Tourism.We have MASS unemployment.Our Foodbanks are empty! NO WAY are these local Welshmen & women!Their 'OUTSIDERS!'..ASK THE WELSH!(NOT the English masquerading as Welsh!) [WELSH FLAG][UK FLAG]
- P3: I expect the BBC would have called the Britains Home Guard 'right wing extremists' in 1940. Especially when they arrested Nazis invaders who landed here ilegally during the Battle of Britain.

Fig. 18

Example 5 shows a final instance of projection, this time framing hateful attitudes towards migrants and refugees as questions which require active engagement from an assumed audience to fully confirm. The initial utterance posted by P1 is accompanied by a link to a BBC article describing "far-right" extremists protesting an army base being used as temporary asylum. P1's claims begin by describing the "typical" behaviour of BBC reporting labelling those who "don't want wales to be turned into a refugee camp" as "extremist". The hyperbolic suggestion opening this utterance describes the idea that asylum being granted to migrants is a steppingstone toward the entire country of Wales being repurposed as a refugee camp. This is bound up with ideas of nationalism, statehood and the replacement of local or authentic Welsh residents by outside refugees. P1 does not initially identify themselves as someone who would be labelled by the BBC in this way, but deploys stake inoculation to

frame their concern and moral outrage as being benevolent and for Wales as a country not only themselves. The use of the word “we” in the following sentence widens the net of attribution and provides an instance of inclusive in-group labelling, presenting both the speaker and the presumed audience as members of the same at-risk group. The “we” presented here are all of those who may be targeted by the “other” designed in the previous sentence, including both the encroaching migrant population and the BBC institution that is apparently intent on increasing the migrant population and demonising those who would oppose Wales’ reconstitution as a refugee camp.

In phrasing this utterance as a question P1 further removes themselves from responsibility for the judgement produced, instead deputising the reader (who is assumed to be a member of the generated “we” category) into actively engaging in deciding the number of migrants who must be allowed asylum to appease the BBC, or anyone else who would demonise the “we” in-group as extremist. Moreover, the information provided by the initial utterance, which may be the entirety of what creates the ICM for the reader’s decision, does not mention the asylum status of the migrants, which may justify or validate their need to enter Wales. They are referred to only as “economic migrants” who would turn Wales into a “refugee camp”, invalidating their motivation and providing a specific, and wide-reaching threat against the “we” group. The out-group is designed in this ICM not as fleeing their home country for safety, but as seeking a rich state to take advantage of. With an absence of countering contextual information, this group designation reduces any moral or ethical issues the reader may have for not condemning the migrants. In this model the people who are threatening to turn Wales into a refugee camp are doing so in search of money, not in search of safety or a better life.

Additionally, at no point does P1 identify themselves as threatened, afraid or impacted by these existential threats. These are threats to Wales and its citizens and the “we” in-group. The individual stakes of this apparent danger are inoculated and instead framed entirely as a community issue, where the local, valid citizens of Wales, their homes and their norms are under attack. Furthermore, any individual from that community who would dare to oppose this danger are labelled by the politically motivated news source as extremist. Not only is this prejudiced hate speech projected upon the reader, but the apparently unjust censorship and

demonisation of those who would produce such hate speech is used as a bolstering justification for its existence. The intertextual gap between the clear invocation of the hate speech genre and the individual producing that genre is manipulated to allow the speaker to avoid responsibility and reframe their speech genre as truth telling and not bigotry.

## Summary

In this chapter, three different linguistic strategies were identified as being used 'in the wild' to manipulate the intertextual gap around the hate speech genre, to be able to produce and achieve its communicative aims while reducing personal responsibility and reframing the genre as one that is more acceptable and importantly acts to avoid platform censorship. In engaging with these different manipulations, hate speakers are in certain instances able to achieve their aims alongside the direct use of slurs, tropes and othering language, by justifying and validating their actions and providing themselves with deniability, where they to be reprimanded.

"Legitimation" was discussed as a technique whereby hate speakers either explicitly or implicitly legitimize their hateful speech through the invocation of authority or creative framing. Examples discussed show how hate speakers can present themselves as objective and indifferent to the hateful rhetoric they produce by framing it as intellectually justified debate, or as imbued with the authority to speak on such a topic because of traits inherent in their identity.

"Excusing" illustrates how hate speakers, aware of the social stigma of broadcasting racist sentiments, combine with their hate speech a pre-designed excuse as a means to discount the expected criticism they assume they will receive. By including efforts towards consensus raising, equating acceptable narratives with unacceptable ones or invitations to the application of "common sense", excuses are made as a preamble (or a concurrent amble) to head off criticism as it may occur, justifying their speech.

Finally, "Projection" and directive speech are linguistic turns which transplant the responsibility of hateful content upon the reader or presumed audience, by way of structure

and framing. In presenting hate as questions, worded specifically for an outside individual to confirm, hate speakers present themselves as distinct from and indifferent to their rhetoric, projecting it outward as a moral decision of the reader. By presenting the othered groups as threats to presumed readers or even the social order of majority communities, the hateful speech is re-framed as a moral good for those not in the identified out-group and the hate speaker themselves is inoculated from any accusation or ulterior motive.

# CHAPTER 5: MEMBERSHIP CATEGORISATION ANALYSIS AND IDENTITY

## Introduction

This chapter will analyse online social media data with the aim of understanding how identity is created, maintained and weaponised in digital interaction. The theory of Symbolic Interactionism, as first conceived by George Herbert Mead (1934) and later developed and nominalised by Herbert Blumer (1969), posits that identity is something one *does* not something one *has.* Identity is achieved, maintained and damaged through interaction. As discussed in the methodology (Chapter 3), it is this theoretical stance that drives this research to focus on the identity work done purely in the text of interaction, forgoing any demographic information that may be available in profiles or avatars. This is not only to redress any assumption that users themselves would investigate further than the text when engaging with another user, but also because of the logistics of the data collection method presenting only text (again discussed at length in the methodology). Beyond the use of online handles and avatar pictures, identities are created through "illocutionary" speech (Butler 1997). That is, speech which performs action, rather that signifies and symbolises it. When an identity, or a category of identity, is invoked online, it *does* illocutionary work to create and maintain that identity. In Goffmanian terms (1959), this performance of online identity, through language and text creates and maintains an identity which is received, interpreted by, and co-constructed with, other social media users in their possible audience. Because of the disruption of the contextual and non-verbal cues that often follow alongside the textual element of discourse in offline interaction, online interaction or "Computer Mediated Communication" (CMC) provides a neat fit for the application of ethnomethodological tools to its analysis.

The research in this chapter will engage a narrow range of those analytic tools, drawing from relevant traditions in discourse analysis, focusing mainly on Ethnomethodology/Conversation Analysis (EMCA), and particularly within this field, Membership Categorisation Analysis (MCA). Continuing on, this chapter will present the data that will be discussed, before moving onto the substantive analysis which will be organised

by recurrent and significant themes found within the data, namely; declarative identity, oppositional identity, and the weaponization of identity.

### *The Data*

Because the analysis in this chapter is structured around reoccurring themes found across the data, that data is presented here in its entirety as a point of reference. Thematically relevant excerpts from the data will be quoted as necessary within the analysis. In each instance the data is presented anonymised and verbatim as collected, and in each instance P1 is the inciting hateful speech, and each subsequent P# is a respondent engaging in counter speech.

- P1: *I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey*
- P2: *I can proudly say that as a human being that I hate pathetic racists like you*
- P3: *You have to give children a little leeway when they say dumb, racist shit.*

Fig. 19

- P1*: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.*
- P2: *hahahahahahahahahaha does that apply to Disney movies? Or making S'mores? What about birthday parties??*
- P3: *I mean, #asshats just make everything stupider. Don't even try and deny it.*
- P4: *maybe you're just a pussy…*

Fig. 20

- P1: *[Video] 200 migrants try to break through Croatian border shouting "Allahu Akbar." They also chanted, "No Croatia, Germany," implying that their intended final destination was a more generous welfare state.*
- P2: *If I wore a pair of underwear for a week without washing them would you let me put them over your head and play you a song on a ukulele until you passed out from ecstacy P1*

Fig. 21

- P1: *But why these people getting mad over me saying black people smell like whaaaat!! Not all smell but some do. I'm not racist. It's the truth*
- P2: *You look stinkier than any Black person I know. Also, more ignorant, low class & lacking in basic human decency.*
- P3: *black people work lots of labor intensive jobs. They sweat. Sweat stinks. Try some critical thinking sometime.*

Fig. 4

Fig. 22

With the data presented, this chapter will turn to its first discursive theme of interest; Declarative Identity.

## Declarative identity

The first form of identity construction this chapter will discuss is identity as self-created through declarative language. That is to say, when an individual explicitly offers an identity for themselves through a declaration in their talk. Referring back to Sifianou and Bella (2019), most, if not all, social media platforms enforce or encourage brevity in their posts (Twitter has a word limit, Facebook enlarges text and offers coloured backgrounds for short posts, Reddit and YouTube collapse longer posts, adding an extra step to reveal overly long content). This necessarily constrains the form users are speaking within, emphasising the conciseness needed to effectively accomplish identity work in the face of restricted

demography. Because of this, any stand-alone statement of identity needs to be powerful and easily identifiable to as many people who may see it as possible, leading individuals to deploy what Sacks (1995) referred to as "inference rich categories". Silverman (1998) suggests that during interactions with strangers common opening questions such as "Where are you from?" and "What do you do?" are used specifically to invoke inference rich categories like place of residence and occupation because they hold within them a wealth of inferred and assumed information. Deploying categories such as these in the outset of an argument or discussion can achieve a great deal of identity work without the elongated pre-amble that is restricted by the structure of online interaction. The following is an excerpt from Fig. 19 showing a two-part adjacency pair, which contain inference rich categories, deployed as declarative identity:

- P1: *I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey*
- P2: *I can proudly say that as a human being that I hate pathetic racists like you*

Focussing on P1 and their declaration, the invocation "American" provides the inference rich member ship category of nationality, identifying them geographically and in terms of wider western ideas and geo-politics. Not only this, but the American identity also provides them with what they imply to be the "authority" to speak upon American matters. van Leeuwen (2007) calls this process "Authorization" and suggests that it provides an interactive member with a way to provide "legitimation by reference to the authority of tradition, custom and law" (p.92) to their utterances. The entire opening clause "I can proudly say as an american" works additionally as a hedging tool to identify the speaker as an in-group member of the community upon which they will pass judgement in the remainder of their utterance. The adverb "proudly" additionally works as a descriptive modifier for the identity as "American", conjuring the archetypal image of the "proud American" and the assumed 'American Exceptionalism' (Koh, 2003) that is bound with that image. This kind of Authorization is most closely identified by van Leeuwen as "The Authority of Tradition" (2007: 96), a form of authority that is commanded "not 'because it is compulsory', but 'because this is what we always do' or 'because this is what we have always done'".

This membership as “proud American” places this constructed identity into two distinct and powerful hierarchies in the presumed understanding of the recipient, American exceptionalism in the hierarchy of nationality, and “proud” American in the internal hierarchy of “Americans”. As Koh suggests, American exceptionalism is a well-known concept, that self-elevates America and its population in terms of perceived importance and achievement on a global scale. Without applying judgement to this phenomenon, American Exceptionalism is something often invoked, for good or ill, in political speeches and by American citizens in the description of themselves of their country, to illuminate their prioritised status. This is a powerful, inference rich category marker, imbued with meaning that is intended to be viewed by those not from America as a sign of status and authority. “Proud American” however, provides a different category, internal within the umbrella of “American” that implies authenticity and legitimacy within that category. This modifier suggests that there are Americans within this category of lower status due to their lack of pride. There is an implicit assumption then, within this declarative identity, that there is a both a prideful and a shameful way to *do* being American, and this individual asserts himself to be doing it the ‘right’ way.

The declaration of “proud” American, in opposition to a shameful variant, can work in Goffman’s (1967) terms as a form of face maintenance, bolstering and legitimizing the speaker’s face in advance of any push back. Because this declarative identity is used as a hedge before a controversial and hateful statement, it can be seen as quick and powerful face work to mitigate any damage to the self that may accompany the hate speech. Additionally, by posturing oneself as *doing* “proud American”, this can produce a weaponised shame that can act as a behavioural policing mechanism to deter others from disagreeing or attempting face threatening rebuttals. Braithewaite (1989) suggests that “sanctions imposed by relatives, friends or a personally relevant collective have more effect on criminal behaviour than sanctions imposed by a remote legal authority” (p.69). While strangers interacting online can obviously not be described as relatives or friends, this particularly strong invocation of national identity provides a useful collective through which to administer a shaming sanction to those who may be accused of performing that identity wrongly. This form of pre-emptive shaming against doing being the in-group wrongly may also be viewed as a form of re-integrative shaming (p.54), providing an explicit avenue for group connection that those who

may have been tempted to diverge can re-join. While there is an implicit threat of stigmatization to those who would oppose the hate speaker's claims, there is an easily discerned group that the presumed audience member may ingratiate themselves into by avoiding any face threatening opposition. Standing against the hate speaker makes you a shameful or wrong American (or even worse, not American at all), but agreement or complicity aligns you squarely within the in-group.

While declarative, "I am…" statements of identity may not be the most common forms of identity construction found in online discussions, it is apparent that meaningful and impactful identities can be constructed with as little as two words understood to be rich with inferential meaning. Looking no further than the two analysed words above, the individual has (or at least has intended to) identified themselves as important and authoritative on a world scale and correct in their internal performance of that authority within his community. This may explain their confidence in using less ambiguous and more proclamatory phrasings like "I can…" to open their utterance and the very harsh and inflammatory language which follows this opening hedge.

Declarative identity work as seen in this example constructs identity in a uni-directional way, emphatically proclaiming identity with no requirement of response or confirmation from a recipient. This is one stark difference between the interactionist analysis of identity online and the face-to-face interaction more traditionally investigated. Online social media interaction is a-synchronous and often without a specific audience or recipient in mind. Depending on the platform, the possible audience of an utterance could range from verified friends and family, to strangers who follow you, to strangers who do not but have had your message broadcast through re-blogging (where one user shares something they have found to their own audience). However, there is still an opportunity and assumption of response, distinguishing it also from other front facing public discourses like speeches or advertisements. Social media discourses are often uniquely treated as uni-directional public discourse, which then later take on the features of face-to-face interpersonal interactions. Fairclough's (1994) discussion of the "conversationalization of public discourse and the authority of the consumer" touches on these ideas, but transferred to social media discourses, it becomes apparent that public discourses are now accomplished with the

conversational practices of the everyday, but interestingly and conversely, what previously would be considered everyday conversational discourses (making your opinions known on politics, tv shows, the weather etc.) are in some ways treated as public discourses, presented in proclamatory style with a broad and anonymous audience in mind.

With this considered, declarative, self-constructed identity may not be as powerful and impactful as those doing the work assume. Although, as stated before, the claimed identity of "proud American" is bound up with implied status and exceptionalism, if that generated archetypal imagery does not cohere with the content of what follows, that identity may only be accomplished weakly, if at all. To (ironically) apply the adage "actions speak louder than words" to this, and any, invocation of hate speech, the social "action" that is linguistically accomplished (blatant and explicit racism) does not cohere with the identity stated. The performed identity does not confirm what was asserted in the proclaimed identity. The impact in the received inference of these two conflicting messages is made apparent in the reply where an identity (differing to the proclaimed one) is received and given back to the original poster; that identity being 'a racist' and not 'a proud American'.

The dominant social action that is accomplished in this initial utterance is the deployment of hate speech and illocutionary (Butler, 1997) violence towards a stated black person, President Obama. In deploying a torrent of dehumanising, racist labels, the hate speaker re-creates the identity of Barack Obama not as the president, or an American or a person, but as a subjugated racial stereotype, performing damage and violence to not only that stated individual but anyone who may feel represented by the racial traits being targeted. Moreover, this violent description puts Obama and any who identify with him explicitly at odds with the "proud" American identity that was performed in the opening to that utterance, compounding the illocutionary, linguistic violence being done. The power of the stereotypes and slurs being deployed in this rhetoric are strong enough that they may, in the eyes of some, undermine the ideals of the identity they have claimed in the opening of their utterance. Whether or not the declared identity works for all who may view this utterance, the illocutionary power of the dehumanizing speech used can still be understood as doing real damage that would be identifiable by viewers and intervenable by the standards set out by both the platform and real-world policing institutions. The respondent counter

speech demonstrates a failure of the declarative identity to provide legitimate authority to such forms of speech.

Anita Pomerantz (1984) explains the concept of Preference Organisation as a feature of the second pair parts in dialogic interaction. Although this form of digital discourse is a-synchronous, there is still an assumed audience that can, if it so chooses, respond and engage with a stated utterance. It is the second part in this dialogic pair that confirms and understands the prior turns talk, reifying the accomplishment of that talk. Pomerantz suggests that members in discourse have a "Preference Organisation" to what they hope that response will be, a preference that is likewise normatively expected. Acceptance or agreement are the normative, expected responses to a statement, otherwise that statement would not be initially made. Rejection or challenge, then, is "dispreferred" and requires an additional response, justification or account to reckon with this interactional disturbance, damage to face, or deviation from the expected line. What is interesting in this case, and in many cases of online hate speech, is that the preferred response is shown within the text of the utterance to not necessarily be the expected response.

The breadth of possible audience to this talk means often it will be received by more than just the intended audience, or the intended audience may be those for whom this kind of speech is skewed from the normative line of public discourse. Marwick and boyd (2010), in their discussion of context collapse on Twitter, note that because of the "various ways people can consume and spread tweets, it is virtually impossible for Twitter users to account for their potential audience, let alone actual readers" and that "without knowing the audience, participants imagine it" (p.119). Twitter, like many other social media platforms, "flattens multiple audiences into one" (p.122), collapsing context for what ever utterance is published and broadcast. Marwick (2013) suggests that this feature stands the current landscape of online interaction in stark contrast to offline life, where it is far easier to "alter self-presentation depending on with whom one is interacting" (p.360). Moreover, through the use of sentinel sites, as were used to gather the analysed data, an even greater number of audiences and audience members are available to view an utterance created with an imagined audience and particular context. In Nissenbaum's (2010) discussion of privacy and technology, she insightfully proposes that "Generally, people are unnerved to discover they

are “known” when they enter what they believe to be a new setting… we feel indignant when other know more about us than we want them to, or when they draw unjustified conclusions about us” (p.50). The developing understanding of online privacy, in a time where networks are becoming more connected and identities more intertwined within and between platforms, would suggest an even more influential pressure from context collapse and an amplified consequence for those who do not recognise and react to that pressure.

Marwick and boyd (2010) suggest that social media users employ different tactics to present an identity online that is authentic, but that also attends to the issue of performing that identity to an imagined, and unknowable audience. In using hedges and authoritative identity declaration, as P1 does, they account for the possibility of an unfriendly audience while portraying a ‘real’ or ‘authentic’ opinion. Marwick and boyd (2010) develop their idea of the “Networked Audience” (p.129), in which the audience is not *entirely* imagined, as they do exist and have agency; agency that is explicitly experienced in the responses garnered from counter speakers or other audience members. Social media users, according to Marwick and boyd (2010) “acknowledge concurrent multiple audiences” (p.129) in the construction of their posts through the interplay of previous experiences interacting with their possible audience, and this may provide an explanation for why such clear and explicit hateful language is performed alongside mitigating identity work.

This applied concept draws clear parallels with one of Mead’s (1934) key concepts in the founding of symbolic interactionism; the generalised other. He notes that in the creation of identity, many identities are performed through interaction within different contexts. When that context is uncertain or collapsed (exemplified by writing a book whose audience can never be fully curated) symbolic interactionists would talk of performing for a “generalised other”. As Marwick and boyd (2010) advocate, in online instances there is often historic feedback that can help to advise future utterances and future identity management, but the crucial uncertainty of the scope of social media resigns online utterances to being created ultimately for a generalised, imagined audience.

The following is an excerpt from Fig. 20, showing another memetic and widely discussed (Bonilla-Silva and Forman, 2000; Nowicka, 2018; Carlson, 2016; Stikkers, 2014;

Archakis, Lampropoulou and Tsakona, 2018) example of a declarative identity performed as a hedge to upcoming hate speech; "I'm not racist but…" which assumes a dispreferred rejection of their talk:

- P1*: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.*

P1 in this interaction from the very outset of their utterance provides an account to justify any upcoming rejection or dispreferred response to their talk. Their declarative identity work states themselves as "not racist", which is then immediately followed by a performance that either *is* or *can* be understood as racist and bigoted. P1 here follows this declarative identity with another authorising assertion to the presumed audience that "you cant deny it". This is a clear example of another of van Leeuwen's (2007) authorising concepts; Theoretical Rationalization . van Leeuwen suggests that, through theoretical rationalization:

> "legitimation is grounded, not in whether the action is morally justified or not, nor in whether it is purposeful or effective, but in whether it is founded on some kind of truth, on 'the way things are'. […] But where naturalizations simply state that some practice or action is 'natural', theoretical legitimations provide explicit representations of 'the way things are'." (p.103)

In the above example, the interactant is justifying and accounting for their declared identity as "not racist" by invoking a form of rationalisation not born from evidence, but instead through an abstracted notion of truth that this belief is naturalised, normalised and just "the way things are" (p.103). This axiom is produced twice (taking up a large portion of their entire utterance), providing justification and legitimation to their identity as "not racist" through an entirely abstracted notion of his assertion being a normalised way of thinking. And through this rationalising account, it is assumed the controversial content within their speech will be accepted.

Both of these interactive pairs show that while the acceptance of the racist statement is the preferred response, it is already understood by the hate speaker that it is not the normative response, and so requires an account even before the utterance is responded too. What is particularly interesting here is that both interactants decided to use a declared identity as an account, presumably because of its assumed power and impact than to provide data or evidence as justification for their remarks. Authority through identity, in this digital social situation at least, is seen as a more affecting justification for face breaking performance than supporting evidence or data would be.

### Oppositional Identity: Creating identity and authority through delegitimising, othering and oppositional speech

Constructing an identity through opposition is something seen often in online discussions, and the power and necessity for oppositional identity is discussed extensively in interactionist literature. Benhabib (1996:3) notes that "since every search for identity includes differentiating oneself from what one is not, identity politics is always and necessarily a politics of the creation of difference", suggesting that identity in both its discrete construction and its use for political or combative interaction is defined as much by the discounting of identity features as it does claiming them. When discussing identity and its existence only as relational, Connolly (1991) states that identity "converts difference into otherness in order to secure its own self certainty". This suggests that not only is identity defined by what it is not, there is a necessity to define external identities as "other" to confirm and reify itself. This may indicate why, from nationalities to sports teams, there is such a tendency to demonise and degrade those who perform differing identities to our own. Identity and its opposition become combative as a self-aggrandising and self-confirming performance.

This is clear and apparent in racist hate speech found online, such as P1 in Fig. 19, describing Barrack Obama as "nigger president Obama stupid porch monkey". Obama was President of the United States, a world-wide representative for the country, voted for and confirmed by the population and its institutions as the constitutional figure head of America. However, despite this the hate speaker attempts to de-legitimise Obama's identity as a

"proud American" (someone who *does* being American correctly) by giving identity to him in opposition his own status as a proud American. Using racial slurs, he conjures historical prejudices and supposed racial hierarchies, that seek to degrade and damage the status of the person they are referring to, delegitimising their position through illocutionary violence (Butler 1997). Essentially re-making and re-constructing the identity of Barrack Obama as incompetent, inferior, less American and therefore less legitimate and authoritative, when compared to the speaker.

This oppositional identity creation is also seen recurring in both Fig. 20 and Fig . 22, where the hate speaker identifies themselves in each instance through their subtle opposition to apparent negative traits held by "black people". In both examples, the hate speaker creates their racial identity not by declaring themselves members of one group, but only in distinguishing themselves as not part of the group they have deemed worthy of attack. They do not give themselves an identity, they do not claim "whiteness", but they occupy a privileged position as a standard category, not the dispreffered category of "black people" who they suggest are "scary" and "smell". It is, as Benhabib suggests, the differentiation from that which they are not that creates, or gives room to create, what they are.

### Weaponising Identity

In moving onto analysing Fig. 21, one can observe associated phenomena being used to illustrate a constructed identity, as an alternative to explicitly stating it:

- P1: *[Video] 200 migrants try to break through Croatian border shouting "Allahu Akbar." They also chanted, "No Croatia, Germany," implying that their intended final destination was a more generous welfare state.*
- P2: *If I wore a pair of underwear for a week without washing them would you let me put them over your head and play you a song on a ukulele until you passed out from ecstacy [P1]*

P1 in this example is able to create and deploy an othering identity upon a migrant group that is rich with implicit and explicit moral and value judgements. The group they are identifying are labelled as "migrants", a definitionally neutral word, but one that is bound with fear and racism in the current climate of right-wing populism and anti-immigration sentiment, garnered by events such as Brexit and the presidency of Donald Trump. The migrant group are tacitly constructed as Muslim through the mention of their apparent shouting "Allahu Akbar". While there is no judgement explicitly placed upon this action, its mention serves as a coded dog whistle (Lopez 2014, Albertson 2014), conjuring Islamophobic sentiment and constructing a negative identity upon the group. This is subtly reinforced by the phrasing "break through the Croatian border", amplifying a violent context to their action, rather than using a phrasing like "crossed". Finally, the last sentence of this digital utterance constructs an identity as conniving, greedy and intent on abusing a "more generous welfare state". This bolsters much of the anti-immigration sentiment and rhetoric propagated by those for whom a dog-whistle such as this is intended. In describing this group of migrants in this politically loaded way, P1 seeks to undermine their identity as legitimate migrants seeking sanctuary and safety, and instead remake them as abusive, conspiratorial and other. While this identity work is being weaponised and done *to* the target group, it is also concurrently and surreptitiously aligning the speaker as native, separate from and suspicious of Muslims, and righteously indignant to the idea of greed and fabricated destitution.

As a final example Fig. 22 shows some more clear and interesting examples of declarative identity work, oppositional identity work, and the weaponization of identity as delegitimation.

- P1: *But why these people getting mad over me saying black people smell like whaaaat!! Not all smell but some do. I'm not racist. It's the truth*
- P2: *You look stinkier than any Black person I know. Also, more ignorant, low class & lacking in basic human decency.*
- P3: *black people work lots of labor intensive jobs. They sweat. Sweat stinks. Try some critical thinking sometime.*

In analysing the opening utterance by P1, what is of interest here is the implication that the controversial talk they are engaging in has previously been pronounced to an audience, who received it poorly and provided a negative, dispreffered rejection. Almost the entirety of P1's message reads as a justification for an interaction that occurred previously and was met with disapproval. However, as is evidenced by the proceeding responses, the use of this rhetorical justification and accompanying identity work fails to avoid the category of racist and is met by delegitimising identity work as supplied through counter speech.

P1 here provides multiple levels of identity in their talk. They engage in declarative speech, stating themselves to be "not racist" as has been seen in previous examples. They also declare their statements to be "the truth", implicitly performing by extension that they are themselves, truthful. This is another explicit and declarative form of identity work, claiming virtuousness as an identity feature and making implicit the notion that those who oppose their racist views are in fact liars. Lastly, the opening section of text in this utterance performs shock and confusion over the idea that anyone might take offence to their upcoming racist ideas.

Once again displaying van Leeuwen's notion of "authority by tradition", P1 suggests that what they are saying is an accepted axiom, engrained in historical social knowledge, and to question and challenge this idea would be performative and wrong. Again, these attempts to build an authoritative identity around their statements as a mitigating hedge suggest that the individual is aware that the content of their utterance is out of step with social norms.

## Summary

This chapter has shown how identity works in action to give authority to statements or delegitimise opponents. Identity is constructed, deployed, co-constructed and reified through interaction, and those identities are made, particularly in this digital setting, entirely linguistically. The discourses used by interactants construct identities in very short spaces where much of the physical and historical identity markers found in other interactive situations do not appear.

Declarative identities are often used to self-create authority (van Leeuwen, 2007), adding legitimacy to proclaimed statements, which in the cases above are racist judgements and comments against othered minority groups. Offensive stereotypes and slurs against racial groups is (mostly) normatively agreed upon to be controversial and deviant behaviour, so authorising identity work is engaged in to hedge the possible face threatening (Goffman, 1967) repercussions of those proclamations. While hate speakers would prefer to be met with accepting and confirming responses, it is apparent that they expect at least some measure of challenge and rejection because of the uncertainty of their imagined audience (Marwick and boyd 2010), and so employ what they believe to be legitimating identities to mitigate the outrage they expect to incur.

While declarative identity work is often used to hedge the judgement they pass against their targeted minority group, often much of the identity work that is received by an audience is gained through the implicit otherness in their opposing language. When othering a group and explicating their negative features, the implicit identity that is created is that you stand in difference to those negative identities. Oppositional language and the innate otherness of defining and giving identity to a person or group provides a powerful portrait of identity that is often more readily and more impactfully received by an audience than any declaration of personal identity could be. This may well be because they are perceived as real and tangible social actions, compared to a more obviously tailored and contrived declared identity. So, while declarative identities are used as a legitimating strategy, they often fail to outweigh the implicit identities constructed through their digital social action.

Hate speech always involves the construction and deployment of exaggerated or entirely falsified identities through which a target is denigrated and attacked. Providing a negative identity *to* a person attaches with it an implicit opposition and mirroring of identity features for the person passing the judgement, so then combative identity work can be understood to provide a dual function in reifying the identity of both attacker and victim. Branding president Barrack Obama as “stupid” and a “nigger” must imply that those doing the judgemental branding are in fact not those things. This identity work may be more

impactful to audiences who oppose racism because they are seen as more authentic and as by-products of their social action, than the identities they themselves choose to invoke.

In the conflict between hate speech and counter speech, identity work is found to be a consistent feature in both the content and process of this interaction. Hate speech's focus on the judgement of goodness and badness of one group when compared to another means that the content of each utterance is very specifically wrapped up in the identities held by or given to a vulnerable minority group. It is those identities that are being deployed or attacked to do illocutionary harm (Butler, 1997) to an individual or group while also using the apparent failures of that group identity to bolster one's own identity in comparison. Through the use of violent, denigrating and dehumanizing speech, particularly slurs bound with contemporary and historical context, online hate speech performs real 'illocutionary' damage upon their targets by re-creating their identities as lesser and inferior. Hate speakers are not using perlocutionary speech to invoke or produce some consequence. In digital space with digital linguistic identities, hate speech is enacting the harm directly. Moreover, the process by which a minority group is judged as lesser, performs as a by-product, the hate speaker's identity as better, and there for legitimate.

In an attempt to perform themselves as the authority and moral superior, hate speakers make broad, confident, declarative statements about their identity ("I'm not a racist but…", "As a proud American…") while often simultaneously weaponizing a negative identity against their target. This weaponized identity provides two functions; in the first instance works to delegitimize the target and remake them as something worthy of their criticism, while also providing an oppositional identity for the hate speaker, which can set them as the privileged standard position in comparison. Not only are they the authority they have proclaimed themselves to be, but they must also be an inversion of all the negative traits they have described in their target.

With identity work in digital hate speech discussed, this thesis will move onto a discussion of linguistic politeness, a field which itself is bound up with the influence and consequences of managed identity work.

# CHAPTER 6: POLITENESS IN HATE SPEECH: POLITENESS IN THE IMPOLITE

## Introduction

Through the combination of hate speech definitions found in the literature review chapter of this thesis (Chapter 2) one can summarise that hate speech can be defined as language used to cause or promote illocutionary harm, discrimination or hostility towards a person or group based upon their protected characteristics (gender, sexuality, race, religion etc.). This definition describes a form of language that is definitionally aggressive, combative and impolite. Stigmatizing, demeaning or attacking a person based on immutable traits can never be described as a polite speech act, however traversing internet blogs or parsing through the data collected for this study, many instances arise where implicitly or explicitly hateful rhetoric are decorated, almost camouflaged, in the markers of linguistic politeness described by Brown and Levinson (1978). Utterances whose core rhetorical objective are to harm are presented alongside polite language that act to save the face of the hateful speaker, the presumed audience who they are trying to convince and even occasionally the target of their hate. This section will discuss three different forms of linguistic politeness as found in the data: Markers of solidarity (positive politeness), Redressive action (negative politeness) and going 'off record'.

## Markers of Solidarity

In her discussion of online politeness in transvestite websites, Planchenault (2010) draws on Eckert and McConnell-Ginet's (2003) definition of positive politeness ("showing that you like or empathize with someone, that you include them in your 'we', your 'in-group'" p.135) to focus their analysis on the use of inclusive pronouns to generate solidarity. Brown and Levinson (1978:70) explain that the use of positive politeness minimises possible face threat by subtly assuring that the speaker must want at least some of what those hearing must want, because they are bound by in-group togetherness. The "we" or "us" that is invoked in the use of these pronouns must share some key connecting characteristics, and so an audience member that hears the face threatening act can rest assured that no negative evaluation will be made upon their face, because they are part of the respected in-group.

When hate speech is deployed with solidarity markers such as these, they are an assertion that any assumed audience member who may be observing can agree with (or at least not explicitly disagree with) that hateful rhetoric without being at risk of losing face, because of their in-group membership.

Solidarity Markers in Hate Speech Examples

1. if white people drive around with american flags everyone talks shit and says **were [we're]** racist but when mexicans do it it's "pride"? um no bye
2. I'm not racist but **we** all know niggers are, idc if your white, black, spic or what the fuck ever. You do nigger things, you are a nigger.
3. #briantaylor Just wat the world needs... another fag is wastg airtime thinkg **we** need 2 hear he's a fag.... #ianthorpe
4. gays/lesbians plz understand **we're** not judging u but thats what God says in his word,ony God can judge
5. MARRIAGE= ONE MAN & ONE WOMAN. Enough of these activist judges. FAVORITE if you agree. I know the silent majority out there is with **us**!
6. Here is another unpopular opinion. If you suck cocks then you are gay and **we** should be allowed to call you a faggot. Because you suck cocks.
7. Apple, why do **we** have to bring gay rights into everything? I don't (for the record) hate gays but you're a computer company.. #forreal

Fig. 23

The above collection of utterances are each from different social media users performing either racist or homophobic hate speech. This collection shows a range of severity and explicitness of hate speech, ranging from outright slurs to low level hateful rhetoric, but each is visibly or understandably hateful, and as confirmation each of these online utterances garnered counter speech responses pointing out their offensiveness.

Each instance above shows some example of a hate speaker using the pronouns "we" or "us" to garner solidarity with, and perform politeness to, an assumed audience that they wish to save face with. Although the hate speaker is engaging in patently impolite speech, positive politeness is paid in each of these examples to those they assume are reading, using these markers to identify that the speaker and hearer are the same, and are part of a collective "we". Particularly pertinent here is Planchenault's (2010) focus on the etymology of the empathy inherent in positive politeness "to suffer with" (p.94). The implication of the

solidarity that is proposed between the hate speaker and the assumed audience is that, more importantly than the possible face threat from aligning oneself with bigoted rhetoric, the in-group that is referred to is under a far greater and more imposing face threat from those outside of the group, whom the hate speaker has chosen to demonise. The empathetic, positive politeness used in these collective pronouns invokes a shared suffering of those in the in-group.

Planchenault also notes in her analysis that the inclusive positive politeness of "we" and "us" pronouns are "used by members in order to reinforce the "boundaries" of the community" (p.96) and as a result identify the "they" of the out-group in opposition. Explicit examples of this are seen in some of the utterances above, where the out-group are unambiguously named and judged. In example 1 the "we" to which the speaker is preaching is openly described as "white people [who] drive around with american flags". This is the in-group which the speaker purports to be speaking about and to, and who must share the characteristic that they are persecuted for their actions, while those in the out-group (Mexicans) are not. The face threat that they are liable to receive is stated outright in this utterance, that they will be deemed racist for their actions, but the inclusive positive politeness produced by the solidarity marker signals that the hearer shall bear no face threat from the speaker because they share wants and needs through mutual identity.

Some examples found in the collection above provide justifications to further the face-saving power of their positive politeness. Example 2 ("I'm not racist but **we** all know…"), 4 ("plz understand **we're** not judging u but thats what God says") and 7 ("why do **we** have to bring gay rights into everything? I don't (for the record) hate gays but") afford excuses and justifications, either individual or collective, for their impolite speech that provides a further shield for the faces of those in the in-group. Further to providing outward protection to those assumed to be in the in-group, this use of collective pronouns has an inverted protective effect too. By including an assumed audience, the legitimacy of the impolite speech and the excuses are increased by a shared consensus. This idea of collectivising excuse making links neatly with Mehlman and Snyder (1985) as discussed in the "Genre" analyses of this thesis, as well as Papacharissi's (2004) discussion of online civility and politeness where in Rousseau's (1994) social contract emphasises the need for consensus following. By including the assumed

audience into their face-saving excuses they expand the protection for others to identify themselves with the proposed in-group and fortify the efficacy of those excuse with a proclamation of raised consensus.

### Going "Off-Record"

The second of Brown and Levinson's (1978) politeness strategies that have been found in the data they term going "off record"(p.69). Going on or off record, according to Brown and Levinson, depends on the level of ambiguity that is performed in making a face threatening statement or request. By going "on record" with a speech act that may cause damage to face, the form of speech is seen as less polite but removes any issues of incorrect interpretation. Taking context and relationship into account between the speaker and the hearer, going on record can be useful in ensuring the message that is being delivered or the request being made is clear and explicitly understood by the person receiving it. Brown and Levinson state that when making a request or statement "on record" there is only "one unambiguously attributable intention with which witnesses would concur" (p.69). If the two members of the interaction are intimately familiar with each other, or the speaker is commanding power or hierarchy over the other, politeness may be disposed of and an unambiguous "on record" utterance can be made without worry that face threat needs to be considered. By being unambiguous with their intention, a speaker increases the likelihood that their utterance will be understood, but in cases where that utterance may threaten face there is no gap of ambiguity through which a speaker may escape from their committed intention. If the face threat of an utterance is understood and rebuked by a hearer, an "on record" statement is difficult or impossible to rescind. In the case of hate speech, by going "on record" a speaker is unequivocally asserting their beliefs and committing themselves to those beliefs with little chance of negating any face threat felt by the target of their attack or any others who might be offended as an audience.

Alternatively, and with increased politeness, a speaker can go "off record" with their possible face threatening speech by speaking ambiguously and alluding to or implying their intention. By speaking in generalities or parallels to their intended topic, they rely on their assumed audience to infer meaning and co-accomplish their face threatening act. However,

in the case that the implied act is received poorly, the "off-record" ambiguity affords a possible "out" to both the speaker and the hearer in allowing them to refuse to interpret the ambiguous "off-record" statement explicitly as what is implied. Brown and Levinson note that in being indirect with the intention of a speech act, that intention becomes "to some degree negotiable" (p.69). That negotiation becomes more essential when the face threatening act conveyed is not reinforced by unequal power dynamics or hierarchy between speaker and hearer. In an online setting, where anonymity and unfamiliarity can be increased, this ambiguous space may be more likely as strangers try to impose possible face threats on other strangers. Off-record racism online can be achieved using "metaphor and irony, rhetorical questions, understatement, tautologies" (p.69) to hint towards their intended speech act without explicit commitment to it.

Going "Off-Record" in Hate Speech Examples

1. It's sad that white on black crime is national news at ESPN. Black on white crime no big deal. #notracist, #cantbejust1way,
2. I'm not racist… But if a black police officer shot a white man nothing would be said. Its not always discrimination. Look at the facts.
3. Look I'm not racist, but this ferguson bullshit is ridiculous. Just because a black man got killed all the black people are rioting
4. Let's project into the future. How would you feel if a gay teacher molests your son in school? Its scary enough having daughters already.
5. Marriage is a relationship between a man and a woman. I don't think it is the role of the state to define what marriage is.
6. [Video]: 200 migrants try to break through Croatian border shouting "Allahu Akbar." They also chanted, "No Croatia, Germany," implying that their intended final destination was a more generous welfare state.

Fig. 24

Each of the above utterances goes "off record" with their hateful messaging by using allusion, rhetoric, metaphor, and understatement to cloak their attacks with politeness. The speakers of each of these messages requires the reader to discern the hateful narrative underpinning their utterance for it to be fully manifest, but leave themselves and their possible audience with room to negotiate away from any face damage that may be felt.

Beginning with the first three examples above, each of them allude to anti-black racism by using hypothetical rhetoric and metaphorical speculation to invoke “polite discussion” about perceived inequality favouring black people over white people. The first example supposes the spurious tautology that “white on black crime is national news” while “black on white crime [is] no big deal”. The off-record intention here is to suggest that black people are favoured, and their offenses are ignored, while white people are overly policed and publicly shamed for the same actions. Whether this is true or not, this indirect and ambiguous way of stating the perceived imbalance allows for the implied message to be received and accepted by the hearer, without the speaker outright stating their intention. Were this utterance to be received poorly both the speaker and the hearer could save face with an alternative interpretation, such as this being a commentary on the reporting practices of biased media conglomerates, as opposed to a discussion of unequal treatment of races. Examples 2 and 3 follow a similar route, where although they are making damaging generalisations about black people, the ambiguity of their phrasing allows for the “out” that they are in fact discussing a lack of equality across races (ex 2) or lambasting uncivil or violent disobedience (ex 3).

Examples 4 and 5 both show homophobic hate speech, but each employ different tactics to paint their dehumanising and victimising speech as polite and civil. Example 4 uses a hypothetical scenario to equate homosexuality with paedophilia, but disguises it as concern for the safety of young children. A rhetorical question is used to provide the hearer with a chance to explain their reaction to a hypothetical attack against a hypothetical son, but at no point is it made explicit that the speaker believes it is the hypothetical teacher’s homosexuality that makes them a violent predator. It is left for the hearer to accomplish the homophobic intent, subtly suggested by the utterance. Again, if this were not received well, there is ambiguous space in which the speaker could recant their messaging and reinterpret it merely as concern for child safety and a proposed thought exercise, not an outright declaration of fear and mistrust against gay people. In a different tactic, example 5 avoids responsibility and going on-record with their homophobic intent by speaking only in generalities to legal definitions. They are not stating their face threatening opinion on gay people and their rights, but alluding to semantic and definitional issues about their right to marry. This could be framed as a discussion of the role of state and religion in marriage, as

opposed to an opposition to the rights of LGBTQ+ people, if face threat was received and argued against.

Finally, example 6 is able to deploy anti-immigrant and Islamophobic hate speech without making any explicit on-record face threats to the chosen groups or those who may be reading. By using understatement and objective sounding journalistic language, the speaker explicates no personal judgements or incivility towards the people they are referring to. There are no overtly negative assessments made against the people identified, and beyond that no identifying labels are given to them beyond the term "migrants". By dispassionately reporting that they were shouting "Allahu Akbar" and that they were seeking a "welfare state", the speaker puts themself off-record enough to profess that they were merely stating facts, were they to be confronted about the implied intent behind their statements. In its totality, this utterance portrays this group as a large, dangerous, ideologically driven, religious fundamentalist and illegitimate group of economic migrants, but if that implication was made explicit by anyone receiving this message, the understated, dispassionate, polite language used can provide the speaker with a means of protecting their face.

### Redressive Action

The final form of linguistic politeness that will be discussed in this chapter is redressive action, a form of negative politeness where face is given to the possible victim, with a view to "partially satisfying (redressing) H's negative face" (Brown and Levinson 1978:70). The speaker, knowing that their upcoming speech act will threaten or damage the face or the hearer, will provide redressive action to counter act that face threat, or at least signify that the damage will be minimised in some way. Archetypal forms of redressive action, according to Brown and Levinson are self-effacement, apologies, hedges, impersonalisation and reassurances that the addressee's negative face is unintended and will be avoided.

When applying this idea to hate speech online, clearly impolite speech acts are redressed often by impersonalising and apologising for the upcoming face threat that is about to be deployed. By making exceptions or directing their critique towards an impersonal "other" group, hate speech is able to be used while face damage to that group is symbolically

minimised. Hedges are used to illustrate that the face threat is not intended, or any face threat that is felt is given begrudgingly by the speaker and that they wish to satisfy their negative face with some sort of reassurance that it cannot be avoided, but would be if it could. In the examples shown below, each occasion of hate speech is coupled with polite hedges that redress the illocutionary damage that is done.

Redressive Action in Hate Speech Examples

1. I don't mind gay people just don't be kissing up in public that's disgusting
2. I have nothing against gay people. I have gay friends&prob some gay family. But I believe marriage should be w. a man&women #itsinthebible
3. Openly gay nigguhs I respect uou more then the down low nigguhs… But y'all look disgusting.. I wanna puke right quick
4. You know some of the senior class are some cool mother fuckers, but god damn most of them are some bitch ass pussy ass niggas
5. I'm sick of being the minority in my own country like I'm sorry if I offend anyone but literally America needs to get their shit right

Fig. 25

The first three examples above each perform homophobic hate speech while hedging their language to minimise the infringement on the hearers "basic want to maintain claims of territory and self-determination" (p.71). That is to suggest that even though an imposition will be made and damage to face is likely, attempts will be made to make those hateful judgements impersonal and indirect. Each of these examples opens with an admission, or a supposition, that despite the speakers upcoming rhetoric they wish no ill will towards gay people. They apply no positive judgement, but the speech they use asserts that the upcoming face threat is accounted for and justified by the lack of direct personal hatred. In example 1 and 3 the speaker makes clear that their issue is not with gay people (who they "don't mind" or "respect") but with visible displays of their sexuality, which they find disgusting. This negative politeness impersonalises these judgements, showing that although face threatening damage will be done, it is only because of the actions of the person and not because of the person themselves.

Similarly, with example 2, this speaker states their lack of problem with gay people, providing evidence by their familial proximity to gay people, but that their desire to marry is wrong. This is not a personal threat to gay people, only to the actions they wish to perform. With these assurances the face threat is demoted from direct attacks on identity to a mere condemnation of their actions. Once again this provides a convenient "out" to the hate speakers if questioned; they don't hate gay people, they have explicitly said as much, they only have a problem with their actions. This sort of framing is reminiscent of the "hate the sinner, love the sin" religious rhetoric discussed by Lomash, Brown and Galupo (2019). This admission of (possibly feigned) acceptance redresses the face threat of condemning visible practices that illuminate homosexuality and present the hateful messaging in a far more polite and safe way. This kind of phrasing also provides an "out" for the target of the attack, should they choose it, to interpret this face threat not as hatred towards themselves, but a disagreement in behavioural choices.

Example 4 performs similar politeness, hedging the upcoming face threat by heaping praise on *some* within the targeted group. For clarification, in the thread that this utterance is pulled from, it is confirmed that this speaker is white and using the word "nigga" to address and judge the black people in the class. As with the previous examples of homophobia, the speaker in this case impersonalises their face threat by positively judging certain members of the group and implying that their upcoming condemnation cannot be race motivated because of these exceptions. This politeness before the impolite once again provides an "out" for both the speaker and the hearer; "I'm not talking about *you,* I mean the others"/ "They aren't talking about *me*, they mean the others".

Example 5 shows a clear example of self-effacement as a strategy for politeness, minimising the perceived face threat by making an explicit apology, seemingly placing the blame for their attack on non-white people on their own feelings, or the failings of the country that 'belongs' to the speaker, and not to the group they are identifying. The messaging of this utterance is an offensive attack on those in the perceived minority i.e. non-white people. The speaker themself confirms the possibility for face threatening offense by stating and apologising for it within the utterance. The apology works to redress the negative face given to the targeted group and the effacement of "America need[ing] to get its shit right", while

identifying it as “my country” the speaker minimises the damage done in their attack on the “others” that make them a minority. Again, there is an admission in this attack on immigrants (and “others” who do not fit the American archetype the speaker claims to be a part of), but the polite phrasing allows an “out” for the speaker and hearer to save face. An alternative interpretation to this speech act could suggest that the speaker is merely raising concerns about immigration policy, or tradition and national heritage, and it requires the hearer to co-construct the racist anti-immigrant narrative that is politely presented.

In each of these forms of politeness hateful speech is reframed with linguistic strategies that help to save the face of the speaker, the victim and/or the possible audience found in online platforms. Positive politeness and solidarity markers are used to identify inclusion between the speaker and those who may receive face damage from their speech, be they the victim or a wayside audience. In deploying inclusive markers, the hate speaker demonstrates similarity and shared values with those reading their remarks, softening their edges and raising consensus in their attribution. By going “off-record” hate speakers make subtle (and not-so-subtle) allusions and parallel reformulations of the hateful face attacks that are coded within their polite language. By using metaphor, rhetorical questioning and dispassionate understatement, hate speakers give space for interpretation and escape from a face threatening interaction where the power structure between interactants is unknown or level enough for those face threats to need attending to. And finally, the negative politeness of Redressive action attempts to satisfy the issue of negative face through hedging and self-effacement. In executing these strategies, the hate speaker identifies that they understand the negative face of their victim or targeted group and show that they wish to minimise any imposition or infringement if and when they can. By impersonalising attacks, the hate speakers use this negative politeness to create “outs” for themselves or others through a more civil interpretation of what has been said.

## Summary

This chapter has discussed ways in which linguistic politeness is weaponised by online hate speakers to make more effective and more acceptable the rhetoric used to achieve their

aims. Hate speech, as seen in the above examples, is strategically bound up with the language form and content of politeness as a means of softening their message, justifying or explaining their rhetoric or providing a logical "out" where the face threat of their hate is received poorly. Because of the potentially unlimited viewership of online utterances, hate speakers may take it upon themselves to cloak their hate in politeness to add protection against backlash, but also to make more appealing their ideology to those who may agree or may be swayed to agree by their rhetoric. By implementing the strategies discussed above, hate speakers may more effectively convince those reading, and those interacting, that their messages are not extreme and hateful, but civil discussions or unfortunate misunderstandings. Using Brown and Levinson's (1978) work on politeness, this chapter has shown examples of hate speakers taking the uncertain power dynamic of online discussion into account, to address the imposition of face threatening hate speech and hedge that inconvenience with politeness strategies.

# FINDINGS PART II:

## Counter Speech

Content Warning: This thesis discusses sensitive topics, particularly racism and homophobia. Please be aware that extreme and offensive racial and homophobic slurs will be presented and discussed within the upcoming chapters of this research. For the posterity of the data and its forensic analysis, examples of these hateful rhetorics are presented uncensored and in full as they were originally collected.

- Counter Speech
  - Chapter 7: Genre and Intertextuality
    - Joke/Comedy
    - Mimicry
    - Reductio Ad Absurdum
  - Chapter 8: Membership Categorisation Analysis and Identity
    - Declarative Identity
    - Oppositional Identity
    - Weaponising Identity
  - Chapter 9: Impoliteness
    - Sarcasm/Mock Politeness
    - Positive Impoliteness
    - Impolite Beliefs

Fig. 26

# CHAPTER 7: GENRE AND INTERTEXTUALITY

## Introduction

This chapter analyses counter speech as a genre of online discourse, unpacking the various definitions explored previously, and using the collected data to show *how* counter speech is actually done by online social media users. In synthesising the many definitions used in previous research, counter speech can be understood as defined by the multiple communicative aims it seeks to achieve, as well as its expected linguistic strategies. This chapter will analyse instances of naturally occurring counter speech, to discern the ways speakers manipulate the intertextual gap between generic and actual speech. In doing so, this chapter will uncover the tactics used to undermine hate speech and inoculate possible audiences (Ede and Lunsford, 1984) from its influence.

The three techniques discussed in this chapter are Joke/comedy, mimicry and reductio ad absurdum. Each of these techniques will be discussed in terms of the linguistic power they have to create effective counter speech, which meet its communicative aims without necessitating the use of counter narratives or requests for information. Counter narratives and requests for information, while being the most commonly proposed forms of counter speech, require a level of background knowledge to be generated and deployed. In exploring the three techniques previously listed, this chapter aims to demonstrate alternative methods speakers can (and do) use to meet the communicative aims of counter speech, while simultaneously generating a more amenable "play frame" (Coates, 2007) in which other counter speakers may contribute. In removing the barriers of knowledge and engendering community activism, these techniques may work to increase the universality that counter speech is often praised for.

## Joke/comedy

Jokes and comedy are found frequently in the replies to hateful speakers as a form of counter speech that avoids the preferred or expected tact of directly challenging the

information or hateful content of the speech. Although this will be the prime focus for understanding "jokes" in this chapter, it is worth noting the wealth of analysis on humour as a tool for justifying or softening the broadcast of hateful language (Billig, 2005), and a releaser of prejudice, where exposure to jokes which normalise the degradation of certain groups leads to a deregulation of prejudiced attitudes towards those groups (Ford et al., 2008; Ford et al., 2014). While Williams (2019) suggests that insults can cause an inflamed response from hate speakers, research into comedic insults suggest that they can have a reintegrative, pacifying effect on the target (in this case a hate speaker). This can lead to a bonding and engendering solidarity in the temporary group audience, which may in turn lead to an increase in unique counter speech responses, something which Williams reports as being effective in the stemming of hate speech. Coates (2007) describes the necessity of co-construction in humour, and suggests that in deploying jokes "play frames" (p.32) are generated, which set the boundaries for collaboration and open the space for contribution. Making jokes at the expense of a hate speaker could then be seen as attempting to re-orient the hateful or serious setting into one where undermining contributions can be made by others.

Continuing with the idea of group bonding, Terrion and Ashforth (2002) discuss the development of solidarity within *temporary* groups, as often found in online settings. Terrion and Ashforth orient their discussion around Meyerson et al.'s (1996) definition of temporary groups:

> "[They] have a finite life span, form around a shared and relatively clear goal or purpose, and their success depends on a tight and coordinated coupling of the activity" (p.167)

While this research was focused on a business organisation setting, these parameters can be applied to counter speakers responding to hate speech; they are finite in length, form around the shared goal of opposing hate speech, and success, as noted by Williams, can be dependent on the co-ordination of many hate speakers engaging with the activity. Comedic insults, or "putdowns", are discussed by Terrion and Ashforth (2002) as having two possible, seemingly conflicting, powers; an ability to quickly form solidarity within a temporary group,

and a possibility to be disintegrative between the joker and the target. In the case of counter speakers coalescing against a hate speaker, these two effects could well be occurring in tandem. They concluded that humour, and particularly put down humour, played a large part in "melding individuals into a group" (p.84) and "fostered a sense of…togetherness" (p.85).

While the illocutionary violence of insults run the risk of upsetting and inflaming a hate speaker, they simultaneously broadcast an undermining of that speaker, engendering feelings of togetherness and group membership with others viewing the interaction and agreeing with the counter speaker. Deploying comedic insults can be seen as stretching the intertextual gap with the counter speech genre; the joke used may be irrelevant to the discussion, or provide no alternative narrative or context to the hateful content, but may undermine the *character* of the hate speaker and empower other would be counter speakers (through engendered temporary group membership) to engage. This empowerment to engage raises the number of possible speakers by reducing the barriers to entry an audience member may feel.

### *Joke/Comedy Data*

Joke/Comedy Example 1

- P1: *Video: 200 migrants try to break through Croatian border shouting "Allahu Akbar." They also chanted, "No Croatia, Germany," implying that their intended final destination was a more generous welfare state.*
- P2: *If I wore a pair of underwear for a week without washing them would you let me put them over your head and play you a song on a ukulele until you passed out from ecstacy [Name]*
- P3: Oh no, this sounds really scary. You've convinced me, I'm going to go buy a bunch of guns now and vote for Trump. Lol, you guys.

Fig. 27

Example 1 shows P2 engaging in an absurdist comedic insult against P1's anti-immigrant hate speech, designing and deploying an irrelevant offer to the hate speaker that suggests the speaker would passionately enjoy dirty underwear and ukulele music. This

counter speech makes no mention whatsoever of the content of the hate speech, opting not to engage in a discussion about the information and its efficacy, but to undermine the authority of the hate speaker. By focusing only on discussing the hate speakers supposed predilections for underwear and ukulele, the counter speaker may be subtly implying that the content of their hate speech is pointless, puerile and not worth engaging with seriously. This may be understood as engaging in what Attardo (2000) calls "relevant inappropriateness" (p.793), where a statement is understood to be sarcastic or ironic because it appears to be so obviously irrelevant to the context. The speaker is not genuinely interested in finding out the answer to their question, they are dismissing the hate speech by being relevantly inappropriate in this setting. This counter speaker is working within the intertextual gap, not deploying any expected counter speech techniques, but reaching to the pre-understood shared communicative aims of the counter speech genre to achieve their goals.

P3 in this example also uses hyperbole and jokes to undermine the hate speaker, without engaging in any discussion of the hate speakers' content. Neither counter speaker (P2 or P3) challenge the hate speakers' equivocation of migrants to the Muslim faith, or that they are migrating to take financial advantage of other countries. Instead P3 chooses to provide an exaggerated response to mock the fear mongering attempted by P1. By ending their response with the phrase "Lol, you guys" the counter speaker emphasises the comedic, joke nature of their message, ensuring that their hyperbolic statement is read as sarcastic and not genuine. Both counter speakers undermine the hateful initial message without offering a counter argument or opposing information.

Joke/Comedy Example 2

- P1: Fuck you N us Obama care u fucking nigger I'll Never do it
- P2: Racists and punctuation are never found together.
- P3: The sad thing is, stupid can often be controlled with diet & exercise, but so few do.

Fig. 28

Example 2 shows two more instances of comedic insults being used to diminish the impact of hate speech without directly challenging its content. P1 employs very explicit racist hate speech in the form of racial slurs while referencing “Obama care”, a U.S. policy that attempted to provide affordable healthcare to American citizens. Rather than address the slur, or the hate speaker’s apparent opposition to changes in health care provision, P2 flippantly connects the hate speaker’s racism with their poor punctuation and grammar, suggesting that racism and good language skills are mutually exclusive. This equivalence implies that racism and poor grammar are connected, and demonstrative of a lack of intelligence on the part of the hate speaker, which undermines them. Equating racism with ignorance in this way may broadcast to any audience that the hate speech that they are seeing originates from an ignorant and uneducated source and is not to be taken seriously. This is achieved without providing an alternative narrative to the hateful content that they imply is incorrect.

P3 performs a similar joke insult, suggesting that P1 is stupid, and that this stupidity can be avoided through simple healthy lifestyle choices. This broadcasts the idea that the hate speaker’s ignorance can be overcome easily, and that their ignorance is a result of their own lazy choices, again undermining their hateful content. The reply’s comedic tone is framed by the sarcastic implication that the counter speaker cares about the hate speaker’s stupidity, denoted by the phrases “the sad thing is” and “but so few do”. Both responses (P2 and P3) avoid a direct challenge to the racist and hateful content of the original comment, the expected or prescribed model of counter speech, but work within the intertextual gap of understood counter speech goals to achieve their aims.

Joke/Comedy Example 3

- P1: Statist/leftist Black Americans: Face it, Obama is a fucking nigger…i.e., a traitor and an embarrassment to your race.
- P2: like your mother was for not swallowing you
- P3: posting this online for everyone to see was a really really good idea.
- P4: I just wonder at what point you looked at this tweet and thought "yep, this'll win 'em over!"

Fig. 29

Example 3 shows hate speech denoted by aggressive use of a slur and political judgement. Neither aspect of this content are addressed in the proceeding responses, but each speaker is identifiably engaging with counter speech as a means to undermine through joke and insult. P2 responds with a crude remark about the hate speakers' mother, implying that the judgements made by P1 about President Obama (embarrassment, disappointment) are shared by their own mother for their choice not to 'intervene' in P1's birth. By referencing P1's mother "swallowing" them, rather than any other way of suggesting the speaker should not have been born, P2 illuminates the comedic tone, using exaggerated and crude imagery. Referring to an opponent's mother in a derogatory fashion is also a common trope in ritual insulting, often called "playing the dozens" (Labov 1997:472). Again, this response does not make mention of the content of the hate speech and challenges none of the political statements or slurs, but in deploying comedic insults achieves the reduction of hate speech impact through an undermining of the hate speaker.

P3 and P4 deploy a similar kind of dismissive joke against that hate speaker, using sarcasm and caricaturing to poke fun at the speaker's decision to post their hateful message. By over explaining what P1 did ("posting this online for everyone to see") and overstating what a good idea it was ("really really"), P3 conveys a sarcastic tone, alerting any reader that they in fact believe the opposite of what they are saying. Without addressing what the hate speaker is actually saying, they imply that it was not worth saying or that their decision to say it was wrong. P4 achieves the same thing by explicitly asking the question of their decision-making process. Neither response makes an overtly negative judgement of the content or

their decision to speak hatefully, but through their tone they achieve a dismissive undermining of the hate speaker and their content.

Joke/Comedy Example 4

- P1: Call me racist if you want. People like Obama, no matter what their color, are the reason for the word nigger to begin with. WORTHLESS POS!!
- P2: great tweet glad you're a real person [P1]

Fig. 30

Example 4 again shows clear hate speech, using a racial slur and a political target, but P2's response references neither. Instead P2 uses sarcasm to dismiss the message and bring into discussion the hate speakers' identity as a real person. The compliment "glad you're a real person", as well as the lack of punctuation throughout, is robotic and oddly phrased. This invokes a mocking tone, which inverts the first part of the response and illuminates a sarcastic distrust in the hate speaker's status as a "real person". Undermining the hate speaker's authentic identity in this way weakens the authority with which they make their statements, announcing to other viewers that the hate speaker may be inauthentic or even a fake "bot" account. Flippant questioning of identity like this acts to undermine the power of the hate speaker, and achieves counter speech aims without engaging directly with the hateful content being produced. This can in turn signal to other would-be counter speakers that a working knowledge of the topic of discussion is not necessary to effectively engage with countering hateful rhetoric.

Joke/Comedy Example 5

- P1: Marriage is a relationship between a man and a woman. I don't think it is the role of the state to define what marriage is.
- P2: Everyone on this thread is dumber after reading your comment. May God have mercy on your soul.
- P3: Pics or didn't happem

Fig. 31

This final example shows homophobic hate speech, where P1 is opposing the right of non-heterosexual couples to be married, asserting that belief and declaring their opposition to the state having power to define marriage. The two examples of responding counter speech ignore the explicit content of the message (as well as the hypocrisy of stating a personal definition in something and then disagreeing with another institutions power to provide a definition) and instead employ comedic non-sequiturs as a means of displaying dismissal or disapproval of the hate speech.

P2 paraphrases a popular comedy movie "Billy Madison", in which one character is chastised by another by saying "Everyone in this room is now dumber for having listened to it. I award you no points, and may God have mercy on your soul" (Davis, 1996). P2 re-purposes that dialogue to disparage the hate speech as being ignorant to the point of damaging the intelligence of those who read it. The second sentence ("May God have mercy on your soul") is used hyperbolically, invoking divine mercy to allow for such ignorance. The direct reference to comedic pop-culture and the overly exaggerated nature of the comment displays a comedic tone to the response, and undermines the validity or authority of the hate speech, without directly addressing any of its statements or problems of internal consistency.

P3 reacts in an even more irreverent style, employing a stereotypical, stock response often found on the internet in opposition to outlandish claims. The comedic elements on display here are two-fold; firstly, the short, sharp response with grammatic and spelling errors betrays a dismissive and disinterested tone, and secondly the request for pictures of an

abstract moral judgement is patently absurd. Without engaging at all with the content of the hate speech, this response manages to dismiss the hate speaker's rhetoric as stupid and unworthy of thoughtful engagement, while simultaneously implying a lack of validity to their hateful statement (Attardo, 2000).

In each of these examples counter speakers are seen engaging in identifiable counter speech, undermining the hate speaker and their speech (Ernst et al. 2017), without directly addressing the rhetorical content with challenges or criticisms. Something that may act as a powerful barrier to entry for a would-be counter speaker, is not believing they have the intelligence, information or authority to engage meaningfully with hate speech, but these response examples show that counter speech is achievable numerously and quickly without a reserve of opposing facts and figures. While "high quality" counter speech that engages ideologically with hate is seen as a gold standard for civil interaction, studies like Williams (2019) also suggest that the number of counter speech responses encountered by a hate speaker is of great importance to its efficacy. Counter speech that circumvents ideological discussion in favour of comedic dismissal and undermining of hate speakers, as shown above, can be a tool that removes barriers for a greater volume of counter speech engagement. By manipulating the intertextual gap between what is expected or preferred towards this style of rhetorical action, counter speech may become a much more attainable goal for a greater number of people.

### Mimicry

Mimicry is another rhetorical technique that counter speakers can use which allows them to challenge a hate speaker without using the expected techniques of requesting explanation or providing countering information. By mirroring the style, tone or structure of a hate speakers messages counter speakers are able to use mimicry as a framework to identify issues of logic or consistency, and initiate reflection upon what was uttered. Research by Richardson et al. (2014) into language style matching suggests that these mimicking techniques can engender a connection and alignment between opposing interactants. In re-using the existing structural framework provided by the hate speech, viewers may be able to engage in counter speech without the necessity of fully formed opposing arguments.

Mimicry is also seen to be used in a variable way by different counter speakers, some responding almost entirely in mimicked structure and language, some only mimicking small sections to make a point. The flexibility of this technique lends to its ease of use for unpractised counter speakers. Like Joke/insults before, this method works within the intertextual gap between the expected counter speech techniques and the actual counter speech used in the wild. This allows counter speech aims to be achieved without the added academic weight of being knowledgeable (providing "reasonable and accurate arguments, facts and figures" – Scheib and Preuss, 2016) about the topic under discussion.

Mimicry Example 1

- P1: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.
- P2: I mean, #asshats just make everything stupider. Don't even try and deny it.

Fig. 32

Example 1 shows an extensive mimic of the hate speech produced by P1, where P2 reuses almost the entire structure of the offending message, and 10 out of the 12 words in the response are taken directly from P1. As with the examples in the joke/comedy section, P2 here does not engage with the argument being made by the hate speaker, but instead re-purposes the language and structure provided by the hate speaker to create an alternative response. The thing mimicked in this example are the hedging language (Lakoff, 1973) used by the hate speaker to reduce the personal responsibility they take for their hateful language. In mimicking this, the counter speaker is able to amplify how hollow and insincere this justification may appear to others.

P2 replaces the targeted group (black people) with the insult "asshat", which is implied to be the hate speaker and people like them. It is important to note here that the target group "black people", a protected racial minority group targeted because of the colour of their skin, is replaced by the group "asshats", in invented category group defined by stupidity or

ignorance. While “asshat” is obviously an insult meant to demean the person, its source of harm is not derived from the dehumanisation of a category of people based around indelible, human traits. The counter speaker avoids replacing one kind of hate speech with another, for example deploying sexist, homophobic or ableist categorisation to damage the hate speaker. By making them the target and repeating their hedges, the counter speaker illustrates the lack of effect those hedges have on reducing any pain or impact the attack has. This is all achieved without any direct challenges to the stated belief that black people make things scarier. The counter narrative that what they are saying is stupid, and that their rhetoric is harmful, is all provided in the sarcastic mimic of the hate speaker’s own linguistic structure.

Mimicry Example 2

- P1: I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey
- P2: I can proudly say that as a human being that I hate pathetic racists like you

Fig. 33

Example 2 shows a hate speech/counter speech interaction which deals heavily in identity work (Sacks, 1995; Silverman, 1998; discussed further in the Identity analyses of this thesis, P.112 and P.159), but also illustrates another instance of tightly mimicked structure in counter speech. P2 mimics the structure and hedging language used by the hate speaker, trading out the identity marker (“American”) for their own (“human being”) and swapping the focus of hate from President Obama and black people more generally, to the hate speaker and racists more generally. The hate speech produced here provides no outside motivation (political or individual action) beyond unmitigated racism, and instead merely professes hate for a specific person using violent racial slurs under the justification that their identity as American affords them that right. Without questioning what drives this hate or providing a counter narrative as to why they are wrong, the counter speaker mimics their sentence construction and uses that structure to provide a counter-attack in kind. By directly labelling the hate speaker a racist, this example can be seen as more in line with the definitional understanding of counter speech, providing a more direct challenge with a label normatively

understood to be undesirable. However, employing this structural recontextualising (Richardson et al., 2014) technique still allows counter speakers to achieve counter speech aims, without directly questioning or countering their hateful rhetoric. For example, the counter speaker does not question what is wrong with Obama specifically, or provide information for why their choice of slurs is unacceptable, they simply label them a racist. This still works within the intertextual gap between what is done here to undermine the hate speaker (mimicking their speech structure, invalidating their authoritative identity, etc) and the presentation of alternative narrative or information requesting that is expected in counter speech.

Mimicry Example 3

- P1: I'm officially scared of black guys #notracist #ihavereasonstobescared
- P2: I'm afraid for our youth if you are representative of it #idiot

Fig. 34

Example 3 shows a much looser mimic of hate speech by a counter speaker, conveying the tone, and a small aspect of the structure, to achieve a similar reflection of the hate speaker's rhetoric. P2 mimics the hate speaker by opening their response mirroring an admission of fear with slightly alternative phrasing ("I'm officially scared of…" – "I'm afraid for…"). The substitution of the target of fear, while using a similar structure, greatly changes the character of the admission. Instead of the self-centred, outwardly hateful fear found in P1's remarks, P2 uses a similar style to change the fear into one of sympathetic concern for "our youth". Instead of directly addressing the inherent wrong of racism and stereotyping black people as being worthy of fear, the counter speaker repurposes elements of their opposition's sentence structure to display sympathy and concern that "our youth" may harbour similar, ignorant views. While this is clearly a judgement on the character of the hate speaker, the tone provided is far less negative and violent than that portrayed in the preceding message. This reformulation may be read as more reintegrative (Braithwaite, 1989), either to the hate speaker or more likely the outside audience, than the inciting utterance. The mimicry of this structure is finished by providing a mirroring hashtag at the

end of their response ("#notracist #ihavereasonstobescared" – "#idiot"). Again, rather than providing alternative information, or requesting a justification for their hate, as most counter speech definitions prescribe, this mimicry technique provides a structure by which people can engage in counter speech which sits within the intertextual. The counter speaker is undermining and denouncing the hate speech without providing an alternative narrative or countering argument which they may not have.

Mimicry Example 4

- P1: Not a racist but Muslims come here, buying Subways and refusing to sell ham and bacon? How about you go the fuck home and sell what you want
- P2: Or how about you go the fuck home and make your own damn sandwich with ALL of the ham and bacon?

Fig. 35

Example 4 shows mimicking counter speech where the overall structure is not replicated, but different phrases from the offending message are extracted and repurposed in a new order to create a response. Again, the counter speaker in this interaction does not challenge the content of the hate speech. They do not discuss the franchising choices of Subway sandwich shops, provide information to show that "Muslim" isn't a race or bring up the logical incongruence of telling "Muslims" to "go home", but instead repurposes their words and sentence structure to dismiss their ignorant speech. Half of the words used in P2's response are taken directly from the hate speaker's initial message. The phrase "how about you go the fuck home and…" is recontextualised by the counter speaker. When deployed by the hate speaker, the connotation is that "home" is another country where the hate speaker expects "Muslims" to come from, and that they should return their if they do not wish to sell certain foods. This invocation of food choices or restrictions in this context illustrates the hate speakers' conceptions of "British-ness", or more importantly local cultural tradition against "other", dis-preferred cultural tradition. What is done *here* is to be allowed to buy and eat pork products, what is done *where "Muslims" are from* is to restrict that dietary practice. When used by the counter speaker, that connotation is changed to refer to the hate speakers'

private domicile where they might make their own sandwich. By repurposing the hate speaker's own words, the counter speaker illuminates how ignorant and flawed they are in their original use. They do not engage with an ideological argument, which they may not have the information to back up, but instead use the mimic structure to engage in a semantic one, in which the word "home" takes on new meaning in the mirrored but alternative context. In doing so they undermine and dismiss the hate speaker and their messaging, without having to provide information as to why that message is unacceptable or wrong.

Mimicry Example 5

- P1: Just saw a gay commercial #Yuck #Disgusting #Eww #Gross #Nasty #Sick #HowManyHashtagsCanIComeUpWithToDescribeHowAbhorrentHomosexualityIs
- P2: just saw an inbreeds tweet #disgusting #sick #shameful #idiotic #HowManyHashtagsToDescribeSuchAMoron
- P3: Just saw an idiot tweet #errmygawd #blegh #eeeww #diseased #howmanyletterscaniaddtomakemystupidpointacross

Fig. 36

This concluding example shows a final instance in which mimicry is used to almost entirely recreate the structure of a hateful tweet as a means to portray a dismissive and undermining tone in their counter speech. P2 and P3 both reconstitute the hate speaker's tweet almost verbatim, substituting the specific words causing the hateful violence with words which target the hate speaker and parody their language. All three members interacting use the phrasing "Just saw a(n)…" to identify a target that their subsequent hashtags will pass judgement upon. The hate speaker broadcasts their disgust at seeing a "gay commercial", whereas the two counter speakers proclaim their disgust the hate speaker, who they identify as an "idiot" or an "inbreed". Following the target identification in each tweet, a series of hashtags are used to either condemn the target directly or, as is the case with P3, are filled with nonsense to undermine and ridicule the structure and basic edifice of the initial hate tweet. P2 repeats some of the hate speakers' hashtags unchanged, while both respondents copy the final hashtag structure, beginning a long unbroken series of words with "how many…". P2 uses this final hashtag to insult and condemn the hate speaker as a

"moron", where P3 uses their final hashtag to overtly mock the structure used by the hate speaker, and demean them for publishing "stupid points" on social media. Both counter speakers are able to tightly recreate the structure of the hate speech, but each is able to slightly alter the categorisation of their target, creating counter speech that responds in two different ways. Once again, neither counter speaker provides counter information to the hate speaker but is able to use a mimicked structure to undermine and provide the effects of counter speech within that intertextual gap.

Mimicking, as with the joke/comedy strategy before (Billig, 2001; Terrion and Ashforth, 2002), does not in itself provide the high-quality, information-based counter speech often described or championed in counter speech research (Schieb and Preuss, 2016), but it does provide onlookers with a framework through which to respond to speech they view as hateful. Rather than being stifled by a fear of not being able to provide an adequate alternative narrative, or a breadth of knowledge with which to debate a hate speaker, these techniques can reduce the barrier for participation and provide an easier entry point for would be counter speakers.

### Reductio Ad Absurdum

The last rhetorical technique that will be discussed is "Reductio Ad Absurdum", a logical argumentative philosophy that deploys an argument pushed to its logical absurd extremes to identify its inherent problem. Novaes (2016) notes that an absurd and impossible scenario presented in this fashion can illuminate logical fallacies in an argument and allow for truth to be discerned. With this argument strategy, a counter speaker can expose logical problems with a hate speaker's rhetoric by taking only what they say and pushing it to extreme conclusions. Without necessarily requiring alternative narratives or enquiring for additional information to explain the flaw in the hate speaker's logic or perspective, this technique allows a counter speaker to shine a spotlight on the underlying flaw in a hate speakers' text, magnifying and expanding the premise to absurd lengths. An additional strength of this rhetorical form is that it allows a counter speaker to not only address the moral or logical flaws in hate speech, but also any flaws in their justification or explanation if one is given. Once again, this form of counter speech allows speakers to achieve the goal of

undermining the hate speaker and broadcasting that to a possible audience, without engaging in the expected counter speech techniques of requesting information or providing alternative narratives. Counter speech rhetoric is able to be designed and conjured only from the illogic of what is presented by the hate speaker.

Reductio Ad Absurdum Example 1

- P1: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.
- P2: I guess what your saying is u would feel more comfortable if black people didn't exist? Fascinating #HitlerMuch

Fig. 37

Example 1 shows P2 taking P1's suggestion that "black people make everything scarier" and pushing it to the absurd logical conclusion that it implies. Using only the internal logic of the hate speaker's presented text, the counter speaker is able to show the extreme ends that following this logic would lead to. P2 develops the idea that if black people bring fear and discomfort to everything, as P1 has asserted, then the way to attend to that discomfort is to remove black people entirely. If black people make *everything* scarier, there is nowhere and nothing to remove them to that they would logically not cause discomfort, so the logical but absurd conclusion to this reasoning, as P2 points out, is that P1 must believe that black people shouldn't exist. This is very obviously an extreme and farcical conclusion to draw, but openly displays what following the assertion that "black people make everything scarier" would lead to if it were accepted unchallenged. No counter narrative is presented by the counter speaker, and there is no explicit request for further information, but this response extends the intertextual gap to achieve the counter speech aim of undermining by satirically presenting an extreme reworking of the hateful assertion.

Reductio Ad Absurdum Example 2

- P1: I'm not racist… But, if black Americans want 2 stop getting shot by cops, maybe they should stop committing crime
- P2: Anytime a black person commits any crime they deserve to be shot dead. Got it.

Fig. 38

Example 2 employs the same counter speech tactic, developing the implicit logic of hateful speech and making explicit the extreme conclusions following that logic unincumbered would cause. P1 suggests that to not be shot by police officers, black people should refrain from committing crimes. The implicit logic of this statement is that being shot by the police is an acceptable and equitable response to committing a crime, and that black people in particular should avoid criminal activity if they desire not to experience these repercussions. P2 draws the implications of this statement to extreme conclusions to show their logical flaws, amplifying and explicating its absurdity. By emphasising "anytime" and "any crime", the counter speaker makes obvious how extreme and unjust a ruling this is. Many crimes are non-violent, many are mundane and to suggest that black people should receive a fatal violent punishment for engaging in any of those is patently absurd, however this is what is implied by the logic of the hate speakers' statement. By presenting an extreme and fully realised version of the hate speakers' suggestion, the rhetoric of this counter speech technique illuminates the fundamental issue with such a stance. Again, the counter speaker does not present an alternative narrative or countering information, they only work to make explicit the logical issue by re-asserting a sarcastic extreme as if they were being genuinely considered. They are undermining the hateful rhetoric using only the logical flaws found in the hate speakers' statement, without the use of any additional information from the counter speaker.

Reductio Ad Absurdum Example 3

- P1: Im not racist but the reality is that black people think they're so brave confronting cops, #MikeBrown did this to himself. Not sorry at all
- P2: so ignoring an order is worthy of immediate execution?

Fig. 39

Example 3 presents similar counter speech to the previous example, this time posing the extreme logical conclusion as a question to the hate speaker. P1 uses the specific example of Mike Brown to disparage all black people as being "brave" in confronting police. The suggestion here being that his ill-conceived bravery of confronting a police officer led to the rightful killing of Mike Brown by that police officer. Here, P2 expands slightly on what is presented by the hate speaker, referencing the fatal shooting of Michael Brown that is alluded to by P1 and equating "confronting cops" with "ignoring an order". This elaboration still achieves the same rhetorical manoeuvre as before, using an absurdly exaggerated finality to pick holes in the hate speaker's logic. P1 suggests that the fatal shooting was brought on by Michael Brown himself, and because "black people" as a homogenous group supposedly all act in a similar fashion they too deserve lethal punishment. P2 summarises this suggestion and reformulates it as a question to the hate speaker, asking them to confirm the logical end of their argument, that the defiance of police should result in "immediate execution". The use of the phrase "immediate execution" is also particularly insightful, making plain a fundamental issue with defending the right of police officers to use lethal force as punishment without trial. The use of "execution" invokes a state sanctioned killing that would be decided by trial, a process that was forgone in this instance, and by extension of the hate speaker's logic, should be forgone in any future instance in which "black people" are brave enough to confront the police. Again, without providing an additional narrative the counter speaker is able to discern problems in the logic, provided solely by the hate speaker, and reconstitute them to instigate reflection upon those issues.

Reductio Ad Absurdum Example 4

- P1: Look I'm not racist, but this ferguson bullshit is ridiculous. Just because a black man got killed all the black people are rioting
- P2: Yeah. Rioting. Every. Single. Person. You nailed it.

Fig. 40

The counter speaker in example 4 uses deliberate punctuation and structure to emphasise the outlandish claim made by the hate speaker, to undermine their racism through association with an explication of their illogical statement. P1 asserts that "all black people" are acting as one uniform group, in response to the alluded "ferguson bullshit". The counter speaker here illuminates the absurdity of the hateful claim by re-wording it and emphasising with punctuation. By making the claim so explicit and clear it is easier for a reader to identify the issue and dismiss the claim as faulty. Instead of arguing against the racist suggestion that rioting in response to an unlawful killing of a black man is unjustified, the counter speaker merely highlights the absurdity of suggesting that all of a particular race are acting in exactly the same way. The counter speaker does not question the assertion or argue against it, but instead sarcastically presents the same sentiment in blunt and emphatic terms with an overly congratulatory comment to finish. Once again, the intertextual gap between what is expected in counter speech and what is actually deployed in counter speech is expanded to include an absurd assertion based on the hate speaker's stated rhetoric. In presenting this assertion in such a way, the counter speaker exposes the logical fallacies in their argument, undermines it and broadcasts those problems to others in the audience.

Reductio Ad Absurdum Example 5

- P1: I'm not racist but all this white people can't say the word nigger is such shit. Cant be that offended if you call yourself it
- P2: Are you sure? Sure doesn't seem like it…
- P1: Obviously haha. My nephew is mixed race. I have black people in my family.
- P3: Yikes. If only they knew that you want to be allowed to call them niggers to their faces..

Fig. 41

In this final example, P1 presents a racist assertion as well as an accompanying justification which is pushed to an extreme hypothetical by the counter speaker. The hate speaker initiates their speech by broadcasting their anger at the social norm of white people being restricted from using an extreme racial slur. Their justification for this indignance, they state, is a belief that if a targeted minority group has reclaimed (Ritchie, 2017) and casualised the use of a slur within their group, its use by others, including their historical oppressors, must too have lost its offending power. When questioned by P2 about their opening assertion that they are "not racist but…", P1 provides further justification for the use of the slur in the form of a familial attachment to the race that they are at risk of offending. A speech act that proponents of Membership Categorisation Analysis (MCA, discussed further in Identity chapter) would call "category incumbency" (Smith, 2017:122) by proxy of a supposed relationship. Their membership to the category of "not racist", and their proposition that they are doing "not racist" properly, is supported by their apparent family connection to black people. P3 combines these two pieces of information, that the hate speaker wishes to be able to say the word "nigger" without offending black people and that they have black family, and derive from it the extreme conclusion that the hate speaker must wish to be able to call their family members by that slur without fear of repercussion. While this is an absurdly severe suggestion, it is only reached by following the mathematical logic of the hate speakers' statements. The hate speaker wishes to say "nigger" without offending black people, members of his family are black people, therefore they must want to be allowed to call their

family by that slur without them becoming upset. The counter speaker does not provide information explaining why it would be unacceptable for white people to use that slur, or ask for the hate speaker to explain their desires or their reason for believing their actions would not be offensive, but instead repurpose only the logic that is stated by the hate speaker to make explicit the issues inherent with it.

## Summary

This chapter has presented analysed examples of counter speech as they occur in situ. In exploring the rhetoric used by counter speakers in their response to hateful speech, this chapter has identified three specific techniques found to be used in the wild to achieve the communicative aims of counter speech, despite any logical or logistical barriers that may present themselves. Counter speech, while not policed as an offensive and criminal form of speech, faces barriers which make it less desirable or attainable as a thing to be done by anyone and everyone, an integral feature of the genre. Much of the existing research and literature on counter speech expects those who engage in it effectively to be equipped with knowledge of counter narratives, or the confidence to challenge a hate speaker for a justification without the worry they will suffer embarrassment at the hands of the hate speaker's response. However, the examples found above show social media users, in situ, working within the intertextual gap of counter speech to achieve those aims without the necessity of this additional knowledge.

By employing absurdity, exaggeration, sarcasm (Attardo, 2000) and ritual insulting (Terrion and Ashforth, 2002; Coates, 2007; Haugh 2014, 2016), the "joke/comedy" technique illustrated how, without addressing the specific racist information of hate speech, hate speakers can be undermined. While the jokes made are often, if not always, at the expense of the hate speaker the ritual of engaging in comedic interactions may be able to undercut the stigma of attacking the hate speaker, and if not, joking may foster an accelerated sense of community and purpose among those standing in opposition to the hate speaker. In generating that sense of togetherness among counter speakers and would-be counter speakers, this technique (like the others) shows alternative avenues for counter speech that may encourage others to get involved.

The “Mimicry” technique uses the phrasing and structure of presented hate speech as a template for returning counter speech. Not only does this allow counter speech to be developed on the fly by people without prior knowledge (taking the hateful speech and mirroring it with minimal alterations) it also allows the flaws of the hate speech to become amplified through presenting them in such a similar fashion. By presenting their own arguments back towards them in an altered state, hate speakers are forced to grapple deeper with their own flawed ideas, rather than grapple with new alternative information (Schieb and Preuss, 2016). This is all achievable to counter speakers without presenting a deep and convincing knowledge of opposing information.

Finally, “Reductio ad absurdum” employs exaggeration and sarcasm (Attardo, 2000; Culpeper, 2005) to make apparent the logical flaws within a hate speakers’ argument. As with mimicry, a benefit of this form of counter speech technique is its reliance mainly upon only what is presented by the hate speaker. This technique takes logical flaws and inconsistencies within the presented rhetoric of the hate speaker and expands upon them until that flaw is explicit and obvious. Once again, without presenting any additional information, counter speakers are able to undermine hateful speech by sarcastically presenting back to them a ridiculous by logical conclusion of following the logic implicit within their hateful ideas.

While not exhaustive, this collection of rhetorical techniques show adaptational strategies (Wang et al., 2014) actively employed in situ by social media users to achieve the aims of counter speech in spite of any barriers to participation they may encounter. Each show ways in which users manipulate the intertextual gap, circumventing the prescribed techniques of the counter speech to achieve the communicative objectives of that genre. These are all examples of how counter speech genre is *actually* done in the “real world” where the highlighted obstructions and impediments to performing the preferred, definitional form of counter speech can occur. Counter speech is not as universal and easily achieved a response as is often suggested and hoped (Wright et al., 2017), but the above examples show how it can be accomplished by users who are untrained, and possibly uninformed of whatever information is necessary to bring substantive, narrative criticism (Schieb and Preuss, 2016) to hate speakers. This chapter has shown that not only is working within the intertextual gap of

counter speech a viable option but is one that is actively occurring in online interaction by "real" social media users. Moreover, a deeper understanding of how people *do* counter speech in this fashion may occasion further opportunity for expanding training and developing a more universalised approach to counter speech.

# CHAPTER 8: MEMBERSHIP CATEGORISATION ANALYSIS AND IDENTITY

## Introduction

As with the corresponding chapter covering the application in hate speech, this chapter will discuss identity as understood through Mead (1934) and Blumer's (1969) Symbolic interactionism, and Goffman's notions of face, performance and dramaturgy (1959). Forgoing the demography that may be gleaned from profiles or avatars, this chapter will look at the identity work achieved in interaction by counter speakers opposing online hate speech. The data analysed in this chapter follows the counter speech respondents to the hate speech discussed in the previous identity chapter and will be once again reproduced here in its entirety as a reference point before being analysed in terms of the same recurrent themes; declarative identity, oppositional identity and the weaponization of identity.

### *The Data*

As before, thematically relevant excerpts from the data will be quoted as necessary within the analysis. In each instance the data is presented anonymised and verbatim as collected, and in each instance P1 is the inciting hateful speech, and each subsequent P# is a respondent engaging in counter speech.

- P1: *I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey*
- P2: *I can proudly say that as a human being that I hate pathetic racists like you*
- P3: *You have to give children a little leeway when they say dumb, racist shit.*

Fig. 42

- P1: *I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.*
- P2: *hahahahahahahahahaha does that apply to Disney movies? Or making S'mores? What about birthday parties??*
- P3: *I mean, #asshats just make everything stupider. Don't even try and deny it.*
- P4: *maybe you're just a pussy…*

Fig. 43

- P1: *[Video] 200 migrants try to break through Croatian border shouting "Allahu Akbar." They also chanted, "No Croatia, Germany," implying that their intended final destination was a more generous welfare state.*
- P2: *If I wore a pair of underwear for a week without washing them would you let me put them over your head and play you a song on a ukulele until you passed out from ecstacy P1*

Fig. 44

- P1: *But why these people getting mad over me saying black people smell like whaaaat!! Not all smell but some do. I'm not racist. It's the truth*
- P2: *You look stinkier than any Black person I know. Also, more ignorant, low class & lacking in basic human decency.*
- P3: *black people work lots of labor intensive jobs. They sweat. Sweat stinks. Try some critical thinking sometime.*

Fig. 4

Fig. 45

With the data now displayed, this chapter will turn to its first discursive theme of interest; Declarative Identity.

## Declarative identity

Counter speech is only ever invoked in response to hate speech. It only exists because of the former and without hate speech counter speech would not be necessary. It is because of this oppositional nature that counter speakers engage in declarative identities far more infrequently than hate speakers. The majority of the identity work engaged in by counter speakers, as far as this data is concerned, is focused on questioning and delegitimising the identities created by the hate speaker. Often it is the delegitimisation of hate and the advocacy for vulnerable groups that does the bulk of the work by which a counter speakers' identity is created and maintained. However, the principles of the declarative identity as a concept are not exclusive to use by hate speakers and can be deployed effectively by counter speakers to provide authority and legitimacy to their speech and bolster their rhetorical aims.

This excerpt shows the initial adjacency pair from Fig. 42, featuring P1's hate speech and P2's counter speech:

- P1: *I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey*
- P2: *I can proudly say that as a human being that I hate pathetic racists like you*

In response to the declarative identity made by the hate speaker in Fig. 42, P2 demonstrates their own ability to create and proclaim an identity when they state in opposition that they are speaking "proudly…as a human being". This declarative identity work once again provides authority, this time shifting the source of that authority from the traditions and assumptions of a nation state in hierarchy within the world to an encompassing categorisation as a member of the human race. While this may not combat the opposing authority directly, it deploys a different form of legitimation through conformity (van Leeuwen, 2007: 96). To explain this form of authority, van Leeuwen suggests that:

> "…the answer to the 'why' question is not 'because that's what we always do', but 'because that's what everybody else does', or 'because that's what most people do'.

> The implicit message is, 'Everybody else is doing it, and so should you' or 'Most people are doing it, and so should you'. No further argument." (p.96-97)

By identifying themselves as a member of a larger group (one that includes their interactive combatant), they claim authority for the traditions and moral norms of that group and enforce them upon their opponent. This also affords the speaker the ability to engage in what van Leeuwen calls "Moral Evaluation" (p.97) of the hate speaker and their ideas, as a way to bolster their own oppositional identity. Van Leeuwen suggests that while moral value can be asserted explicitly with words like "good" or "bad", in many or most cases "moral evaluation is linked to specific discourses of moral value" (p.97). Through the deployment of morally charged adjectives, van Leeuwen suggests words like "pathetic" (as P2 uses) trigger a moral concept in a way that avoids being debatable. More explicitly, words like "racist" invoke morality stances that although not universal, are generally looked upon to be negative and morally repugnant. In declaring that they "hate pathetic racists like you", P2 provides examples of oppositional language, that re-constitute the declared identity of the hate speaker, provide moral judgements on their linguistic actions, and re-affirm their own identity as morally righteous and opposed to their opponent.

### Oppositional Identity: Creating identity and authority through delegitimising, othering and oppositional speech

Continuing on with the interaction between hate and counter speakers in Fig. 42., the two adjectives used to describe P1 provide a moral evaluation from the speaker, each providing distinct justifications for the response, while simultaneously legitimising and boosting one another.

The use of the word "hate" provides an abstracted moral evaluation of the described person or action as something worthy of hate, and therefore something offensive and abhorrent. This hatred, which is an extreme form of emotion to place onto someone, is legitimised by the previous identity work gaining authority by conformity with the largest group one can identify themselves with. The use of the word "hate" and its moral evaluations of the hated subject's "badness" imply the counter speaker's innate "goodness" in opposing

it. So here this response is not only providing a differing identity to the hate speaker but bolstering the counter speaker's own declared identity as opposition, delegitimating their opponent and legitimating themselves simultaneously. The word "pathetic" also accomplishes similar moral evaluations of the person and their actions, judging them to be bad, while also implying the opposite features in the counter speaker. This word also provides internal symbiotic justification with "hate", suggesting that the counter speaker hates their opponent because they are pathetic, and that the given identity as pathetic is itself worthy of hate.

The word "racist" in this insulting response also provides for moral evaluation, but is additionally imbedded with a negative identity informed by de-legitimation through theoretical rationalisation, and authority by tradition and conformity. "Racist" is undoubtedly a term full of negative moral and value judgement in its popular use, but it is not by definition an inherently negative term. What saturates negativity into that term, and that identity, are the social traditions by which its negative connotation is given authority, and the social conformity within its community of use that gives consensus to that negativity. So it is then, that to be constructed under the identity of racist is to delegitimise ones person and ones ideas, as they fall outside of the traditional and conforming authority which overrides them. However, this negative identity construction is not explained in the response, but treated as a theoretical rationalisation, that being a racist is bad because that is "the way things are". This combined utterance then accomplishes a duality of identity work, constructing and confirming an identity for the self through opposition to racism and delegitimising their opponent's declarative identity and co-constructing a new one in concert with their racist statements.

Moving on to the following example of hate and counter speech interaction, P2 in Fig. 43 deploys a de-legitimating identity against their interactant partner in an alternative way, not declaring an identity outright, but constructing one through stereotypical associations with the chosen identity, namely a child.

- P1*: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.*

- P2: *hahahahahahahahahaha does that apply to Disney movies? Or making S'mores? What about birthday parties??*

When providing a dispreferred response to P1's hateful idea that "black people make everything scarier", P2 questions whether that assertion holds true in a series of patently absurd situations. P2 here chooses to reject the declarative "not racist" identity declaratively performed by the hate speaker, and instead latch onto the emotive notion of being scared, reconstituting the speaker's identity through association with being fearful and frightened. "Disney movies", "S'mores" and "birthday parties" are all phenomena associated with children, an inference rich age group category that can be deployed as an insult to those who believe themselves to be, and perform themselves online to be, grown adults with educated, well-informed ideas that need to be heard. By re-constructing the person's identity this way, they attempt to severely delegitimate their identity and undermine their proclamations. The identity that was once proclaimed as morally sound and universally agreed ("not racist" and "you cant deny it") has been manipulated and repackaged as one defined by infantile fear and a lack of rationality.

### Weaponising Identity

Moving on to Fig. 44 identity is created and deployed in a new form by both hate and counter speaker. By referencing aspects of disprefered character, each speaker in this interaction powerfully weaponises identity through association to other and damage their chosen targets.

- P1: *[Video] 200 migrants try to break through Croatian border shouting "Allahu Akbar." They also chanted, "No Croatia, Germany," implying that their intended final destination was a more generous welfare state.*
- P2: *If I wore a pair of underwear for a week without washing them would you let me put them over your head and play you a song on a ukulele until you passed out from ecstacy [P1]*

Rather than specifically referencing the implicit racism and anti-immigrant xenophobia in P1's original utterance, it is the performed sense of status and righteousness that the response from the counter speaker (P2) delegitimises with their weaponised identity work. The language used in this response delegitimises the hate speaker by vividly constructing a scenario in which the hate speaker is presented as a sexual deviant not to be taken seriously, opposing the righteous and forthright identity they have tried to perform. The language used here is absurdist and creates an identity entirely unrelated to the one tendered by the hate speaker, and also unrelated to the racist content they delivered. Instead, they weaponise a comical identity as someone who would experience extreme "ecstacy" to the point of unconsciousness at the combination of unwashed underpants and ukulele music. By framing this identity work as a question, it implies a pre-existence of this identity, subtly confirming its reality as constructed by the respondent. All of this works to delegitimise the authority with which P1 imbues his ideas, through their identity as oppositional to 'greedy Muslim migrants'.

P3 and P4 in Fig. 43 achieve a similar de-legitimation in a more deliberate and explicit way:

- P1*: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.*
- P2: *hahahahahahahahahaha does that apply to Disney movies? Or making S'mores? What about birthday parties??*
- P3: *I mean, #asshats just make everything stupider. Don't even try and deny it.*
- P4: *maybe you're just a pussy…*

P4 latches again onto the performance as "scared" and extrapolates that to a full identity summarising and declaring them to be a "pussy". P3 weaponises a different delegitimating identity, undermining the hate speaker by directly attacking their intelligence by proclaiming them to be an "asshat", and that their presence makes things "stupider". In each of these examples of giving identity to an interactive partner, what is seen is a rejection

of the preferred identity that was performed in the initial interactive turn, and a re-construction and deployment of an alternative. This new identity is a co-construction which ignores the preferred identity set out by the hate speaker. Instead, the counter speaker's use the additional racist content that occurs alongside the authoritative identity, to inspire a shameful and de-legitimising identity. This also works as a subtle indication that the declarative identity work they had hoped to achieve, and which was intended to offset and mitigate their racist assertions, has in fact failed. The hateful rhetoric they have deployed has been more effective in associating them with an inference rich membership category (a racist) than their contrary identity declaration.

Both P2 and P3 in Fig. 42 use the newly constructed identities they have created for the hate speaker to not only identify the speaker as unworthy of authority or legitimacy, but also, in tandem set themselves in opposition to those identities:

- P1: *I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey*
- P2: *I can proudly say that as a human being that I hate pathetic racists like you*
- P3: *You have to give children a little leeway when they say dumb, racist shit.*

Both members identify the hate speaker as a racist, setting themselves in opposition and therefore identifying themselves as *not* racist, but P3 goes further to identify their opponent as a member of the group "children", which they, as an implicit adult, must "give leeway to". While this designation as childlike isn't inspired by content from the hate speaker's utterance, it does perform the same linguistic action of delegitimising their speech, and (this time more explicitly) constructing the counter speakers' identity as 'adult', placing them above their opponent in the traditional hierarchy of power and priority. This newly applied identity undercuts their declaration of "Proud American" and replaces it with the imagery of a child throwing a tantrum. Once again, the counter speakers in these interactions choose to avoid engaging with the hate speaker with evidence and counter arguments, and instead embrace the power of shame (Braithwaite, 1989) and identity to delegitimise their racist assertions. The hate speaker has used inference rich (Sacks, 1995) descriptors of virtuous national identity to construct and attempt to maintain their face (Goffman, 1967)

against potential threats, but the counter speaker has instead focused on the violent crude outburst to remake that identity in three ways that are judged to be shameful; as dumb, as racist and as a child. Braithwaite (1989: 71), interestingly, discusses the effect of shaming specifically as it regards to the familial policing of deviance. He notes that the shaming effect of words used to describe deviant behaviours are a useful and necessary maintenance strategy, especially when the logistics of punishment is obfuscated by large populations, and preferred to more purely punitive action which provides no social conditioning to dissuade future deviance. The shaming produced by P3 here not only applies the dishonourable labels to the hate speaker that may work to oppose his messaging, but explicitly invokes a childhood state where the shaming of unwanted and uninformed deviant actions would be necessary, compounding its effect.

As a final example Fig. 45 shows some more clear and interesting examples of declarative identity work, oppositional identity work, and the weaponization of identity as delegitimation.

- P1: *But why these people getting mad over me saying black people smell like whaaaat!! Not all smell but some do. I'm not racist. It's the truth*
- P2: *You look stinkier than any Black person I know. Also, more ignorant, low class & lacking in basic human decency.*
- P3: *black people work lots of labor intensive jobs. They sweat. Sweat stinks. Try some critical thinking sometime.*

Both respondents to P1's hateful assertion diverge from the presented identity put forward by the speaker, and instead construct their own based on the identity feature they most prominently perform; their stupidity. P3 begins their response by rejecting the assertion and providing a rationalising account for where that false notion may have come from. This attempt by the counter speaker to almost justify why their opponent has offered these ideas speaks to Goffman's (1967) idea that the embarrassment from a face breaking misstep in social interaction is felt by the audience as much as it is by the miss-stepper. The "ritual

disequilibrium" (p.304) must be made right to satisfy the ritual state of interaction, and so engaging in the corrective process the interactive partner offers an account for why the out of step statement was made. This correction then also acts as a support to the following weaponised identity that is given to the hate speaker, that being someone incapable of "critical thinking". This again works as a weaponised identity, suggesting that the hate speaker is stupid and incapable to finding the rational account that they did.

The shaming performed by both speakers, but P3 most explicitly, frames the hateful comment as something borne from a lack of understanding, and something that could have been avoided. This links back to Braithwaite's (1989) discussion of on the power of shaming to be a more useful response to deviance than punishment, in that rather than "a denial of confidence in the morality of the offender" shaming can express "personal disappointment that the offender should do something so out of character" (p.72). Although both counter speakers execute their responses in very aggressive and combative ways, much of the shame they attribute to the hate speaker (being ignorant, lacking decency and lacking critical thinking) are things that the hate speaker could achieve or decide to do. In shaming the choice not to engage thoughtfully with their hateful ideas, the counter speakers affirm the ability of the hate speaker to hold the proper morality and the normative ideals of the community. The counter speakers then are doing more than merely applying stigmatising and illocutionary violent labels (Butler, 1997), they are shaming a member of the community who has the capability to 'step' properly within the social norms but in this case has not. This also works as an oppositional identity, illustrating that the counter speaker did use critical thinking and therefore performs the identity of someone who is intelligent, capable and authoritative in their assertions.

P2 opens their retort commenting on the looks of the hate speaker, suggesting that they must have access to a picture of their opponent, most likely through their online profile. They then continue to use this visual analysis to create for the hate speaker a delegitimating identity as "ignorant, low class and lacking in basic human decency". Goffman (1963) discusses this idea, instead coining it a "spoiled identity", where he remarks that:

> “This discrepancy, when known about or apparent, spoils his social identity ; it has the effect of cutting him off from society and from himself so that he stands a discredited person facing an unaccepting world.” (p.30)

These three identity features undercut the hate speaker’s attempt at authority by passing judgement upon their intelligence, their class status and their supposed lack of moral standing. These are three distinct, but related identities, all of which attempt to subvert the “not racist” and “truth telling” identity declared by the hate speaker. They are also, interestingly, created as oppositional identities to a non-specific and abstracted group of “black people”, while also re-affirming P2’s identity as *not* “ignorant, low class and lacking in basic human dignity”. The direct comparison made in the language is between the hate speaker and black people, but the judgement of the hate speaker positions P2 as oppositional and constructs their identity as imbued with authority as intelligent, higher class and decent.

### Summary

When compared to the identity work performed in the hate speech analysis, counter speakers are found to engage much more frequently with oppositional identity and weaponization, while performing comparatively fewer declarative identities.

This may be a direct result of counter speech as a form of language existing purely as oppositional, in that it does not seek to proclaim its own primary statements but at instead seeks to undermine and disprove the statements of others. This may most easily and most readily be achieved through identity work which damages and delegitimises the hate speaker.

In most cases, identities that are weaponised against a hate speaker are based on some part of the performance that the hate speaker has engaged in. Regardless of the declarative identities, the features of their linguistic performance are what holds a much greater impact over the actual identity that is provided and received, by counter speakers at least. The identities they declare may be more useful to and directed towards, members of their same group or opinion. Identities like "not racist", "truth teller", "authority by age/race/nationality/tradition/education etc" may all be convincing and reifying to those who

agree with the upcoming statements and may indeed be taken as 'authority giving' in the right setting and by the right person. But when engaged with by those opposing hateful racists speech, their linguistic actions speak louder than their identifying words. Their identity is constructed by the content of their character, not by the dressings of their digital bodies, and it is those digital dressings that counter speakers most commonly take aim at.

When given restrictions in terms of non-linguistic identity markers, and restrictions for time and space in which to display a self, identity work becomes important and impactful as a way to perform legitimating or delegitimating action. This is shown by the frequency with which interactive partners decide to use identity work as the main focus of their discussion, rather than competing evidence. By delegitimising the character of a hate speaker, counter speakers may undermine their integrity and authority in the eyes of their target audience. This can reduce their influence in general, rather than merely showing them to be ill-informed on the specifics of their chosen topic. The abhorrence of racism and its demonisation is normalised at this point, so the use of engaging with a hate speaker in terms of the logic of their thoughts and the content of their hate, may seem futile. What seems to most likely occur in counter speech as it naturally arises, is weaponised identity work that removes the power of a hate speaker and re-constructs their identity as unworthy of opinion, rather than merely wrong in the opinion they hold.

Whether this is the correct tactic to tackle hate speakers is unknown, but what this chapter has shown is that in a small amount of time and digital space, a great deal of moral meaning can be generated through the identities. The increased fluidity and uncertainty of identity in online spaces makes them more susceptible to re-construction at a moment's notice. Online identity is a constructed performance that must be succinct in both its self-generation and its weaponization against an opponent. With less of the concreate identity markers found in the terrestrial plane (money, status, fashion, profession etc), online identity work and maintenance appears to take on an even more central role in legitimising one's social actions.

With this discussion of identity and counter speech drawn to a close, this thesis will now turn to investigate linguistic impoliteness, a tool whose key aim is damaging and delegitimising the face held in the hate speaker's oppositional identity.

# CHAPTER 9: IMPOLITENESS IN COUNTER SPEECH - IMPOLITENESS IN THE POLITE

## Introduction

As addressed in the genre chapter of this thesis, counter speech is (when compared with hate speech) far less strictly defined because it is not innately a violent or policeable form of speech. Where specific definitions are needed for hate speech to be properly identified and intervened upon by platforms and legal entities, counter speech is a more academic term used to describe "any direct or general response to hateful or harmful speech which seeks to undermine it" (Williams 2019:40). Ernst et al (2017) expand upon this stating that "counter speeches aim at challenging these transmitted ideas of hatred, prejudice or even extremism" (p.7), hinting towards potential strategies for performing counter speech. Possible strategies are further teased out by Schieb and Preuss (2016) when they note that "Reasonable and accurate arguments, facts and figures, employed in direct response to hate posts, are seen as a helpful treatment to restrict the impact of hate speech" (p.5). As discussed earlier in this chapter, polite speech is used to help to save or avoid damage to face in instances where face threat cannot be avoided. When speech is made that can infringe on the agency or identity of a person, politeness can be used to soften that damage or mitigate through escapes from the face threatening situation. Comparing the idea of linguistic politeness to the ideas purported in defining counter speech, obvious parallels present themselves; counter speech undermines and restricts the impact of face threatening hate speech.

However, as was covered earlier, online speech must take into account the faces of three distinct participants, the speaker, the target and the possible hearer. While hate speech is targeted upon an individual or group who is marked out by their at-risk features, the speech can also be face threatening for the speaker who is outing themselves as bigoted and face threatening for a possible hearer who does not identify as part of the target group who may be seen to be agreeing with or condoning that hateful speech. Scheff (1987), Scheff and Retzinger (2001), and Braithwaite (1989) discuss the role of reintegrative shame in their work, a concept that is often discussed in regard to dealing with hate speech, and much of the research and definitional work into counter speech aligns with these ideas. Counter speech

as promoted by reports like Williams' (2019) indeed propose anti-antagonistic counter speech which seeks to provide reintegrative, rather than stigmatising forms of shame in their efforts to undermine and reduce hate speech. However counter speech as it occurs in the wild, performed by the untrained "everyone" that counter speech proponents tout as a key strength, is often combined with a more stigmatising shame that comes from purposefully impolite speech. While the overall aim of counter speech would seem to line up with the concepts informing polite speech, any challenge to a performed narrative will come with a level of face threat. What may be visible in "organic" counter speech though, is the "not unintentional" impoliteness defined by Culpeper (2005:37). In attempting to save or protect the face of a marginalised group, counter speakers can often be seen to be actively impolite to hate speakers, and in some cases actively impolite to others who may identify with some aspect of this impoliteness. This chapter will now, using Culpeper's breakdown of impoliteness strategies, analyse examples found in the data set of speech which achieves the definitional aim of counter speech while actively and intentionally performing linguistic impoliteness.

## Sarcasm/Mock Politeness

### Sarcasm/Mock Politeness in Counter Speech Examples

Eg.1
P1: I mean, I'm not racist but even you cant deny it. Black people make everything scarier. Don't even try to deny it.
P2: not ridiculous at all

Eg.2
P1: But why these people getting mad over me saying black people smell like whaaaat!! Not all smell but some do. I'm not racist. It's the truth
P2: rofl, "not all black people smell" not racist now for sure

Eg.3
P1: OH WAIT I CANT GET IT BECAUSE I DON'T HAVE HEALTH INSURANCE BECAUSE OBAMA IS AN INCOMPETENT NIGGER
P2: Mental health treatment is what you need, right?
P3: https://www.healthcare.gov ¬† here you go bro get some insurance

Eg.4
P1: I can proudly say as an American I fucking hate our nigger president Obama stupid porch monkey
P2: Feel better now that you got that off your chest? It's all just EATIN' you UP, isn't it?

Eg.5
P1: Just saw a gay commercial, #Yuck #Disgusting #Eww #Gross #Nasty #Sick #HowManyHashtagsCanIComeUpWithToDescribeHowAbhorrentHomosexualityIs
P2: Hey, relax. You don't have to have sex with a man if you don't want to.

Eg.6
P1: Not a racist but muslims come here, buying Subways and refusing to sell ham and bacon? How about you go the fuck home and sell what you want
P2: I'm sure you'd say the same about a non-Muslim who opens up a vegetarian restaurant, right?

Fig. 46

The first impoliteness strategy that is exemplified in the excerpts above is Sarcasm/mock politeness, where polite language and politeness strategies that were discussed in the first half of the chapter, are deployed in a way that is obviously insincere. In each instance the example shows the inciting hate speech and the mock polite response used by the counter speaker to undermine the hate speaker's messages and reduce harm.

Examples 1 and 2 both show counter speakers sarcastically stating that the hateful person or language is not “ridiculous” or “racist” at all as a form of insincere positive politeness. In stating that they are *not* ridiculous or racist, they are overstating the act of giving face to the hate speaker, broadcasting that the hate speaker had done something which had indeed damaged their own face. The first example identifies its sarcastic nature by offering a deadpan analysis which seemingly is not required. They are not responding to the messaging or the content of their speech, merely stating without provocation that the hate speaker and their utterance are “not ridiculous”. By performing this mock politeness, the reaction they are causing is to highlight the ridiculousness of the hate speaker’s idea that “black people make everything scarier” and undermine that idea in the eyes of other possible readers. The second example makes their sarcasm even more explicit with the opening “rofl” (rolling on floor laughing), suggesting that they are already treating what the hate speaker said as worth laughing at. By setting the tone with that acronym the counter speakers’ isolation of the hate speakers attempt at hedging their own racism and the analysis that this hedge means that the speaker is “not racist now for sure” are shown to be obviously insincere. Both examples use blunt overstatement and the language of giving face to the hate speaker in such a way that undermines any attempt by the hate speaker to protect their own face using hedges or attempts at justifying their hate speech.

Examples 3, 4 and 5 show mock politeness that appear to emulate redressive action, using negative politeness to address the face damage and minimise its impact. The counter speakers here use sarcastic justifications and solutions to undermine the hatred espoused by the hate speaker by suggesting there is an easy and obvious remedy for their antagonistic stance. This technique minimises the power of the hate speaker by implying that the hate speaker’s wrath is caused by their own problems or confusion, not any features of their targeted group, and that those problems can be attended to easily with the mock redressive action. In example 3, two counter speakers offer seemingly obvious and easy solutions to the problem stated by the hate speaker as their reason for using explicit and aggressive hate speech. P2 references the hate speaker’s anger at not having “health insurance” and uses that to provide what the present as an obvious solution to their problem, that the anger they are performing requires mental health treatment to fix it. By presenting this solution, the implicit sarcastic suggestion is that the hate speech they are using is borne from mental health

issues, not from ignorance and bigotry, and this ironically performs *giving face to* the hate speaker because they now have an excuse for their hate. P3, in an even more casual and flippant manner, sarcastically implies that the rage and racial hate displayed is done to nothing other than the administrative inability of the hate speaker to gain health insurance. By providing the link to the health insurance website, which can plainly be seen as the obvious and simple first step in obtaining insurance, the mock help that the counter speaker provides works to illuminate the stupidity of the hate speaker while satirically providing them with the face-saving 'out' of not knowing what website could provide them with insurance.

Both of these counter speakers offer mock suggestions for the reason the hate speaker is being hateful, undermining any attempt by the hate speaker (or someone who agrees with the hate speaker) to use those reasons as justification for their hateful language. By using the rhetorical "right?" and the casual "here you go bro", each counter speaker establishes a tone of calm, ironic friendship in contrast to the fully capitalised rant and use of slurs. Examples 4 and 5 both show similar mock politeness, where casual and friendly language is used to suggest a simple alternative reason for the aggressive hate speech, which works to discount that reason and place the emphasis squarely on the ignorance and bigotry of the hate speaker. In sarcastically presenting face-saving excuses, this strategy instead confirms and reifies racism and bigotry as the source of their outbursts. P2 in example 4 presents the hate being reacted to as merely a temporary outburst that once expelled will exorcise the hate speaker of the need to use violent racist language and the counter speaker in example 5 mockingly clears up the misunderstanding that seeing a "gay commercial" would contractually oblige the viewer to have sex with someone of the same gender against their will. Both satirically present easy solutions the work to underscore the source of the speaker's hatred as being bigotry against their chosen group.

Finally, example 6 shows an example of mock politeness which utilises a sarcastic form of "off -record" politeness. In this instance the counter speaker is providing an ironic alternative excuse that would have allowed the racist hate speech to be off-record and more distanced, had the hate speaker chosen to use it, and in doing so is removing a possible avenue through which the hate speaker may excuse their speech. As the counter speech in example 3, the rhetorical use of "right?" brings a conversational and provoking tone to the

suggestion and signposts that the attempt at taking the hate speech off-record is insincere and meant to perform and opposite function. By making explicit that the hate speaker could have framed their critique in this way as a cover makes more explicit that the hate speech is intentional and that their target is not chosen at random, increasing face damage to the hate speaker. This impolite politeness by extension reveals this same idea to any other person who may be viewing the hateful rhetoric and attempting to explain it away using that comparison.

Each of these examples shows mock politeness and sarcasm being deployed to illuminate the hate speech that is being presented, remove opportunity for escape and intentionally cause face damage to the hate speaker while attempting to engage in the counter speech aim of reducing harm to those beset. They are achieving a duel aim of reducing face threat and damage to the marginalised target of attack and increasing it to the attacker while using an ironic facsimile of polite language.

### Positive Impoliteness

As mentioned previously, Culpeper's (2005) idea of positive impoliteness is defined as an inversion of positive politeness, where in the impolite person attempts to damage the receiver's positive face wants, those wants being to be "treat[ed]… as a member of an in group, a friend, a person whose wants and personality traits are known and liked" (Brown and Levinson, 1978:70). Culpeper (1996) devised a taxonomy of positive impoliteness strategies including:

- *Ignore, snub the other* – fail to acknowledge the other's presence
- *Exclude the other from an activity*
- *Disassociate from the other* – for example, deny association or common group with the other; avoid sitting together.
- *Be disinterested, unconcerned, unsympathetic* (p.358)

While some of these strategies were devised as physical or prosodic, they can be seen being accomplished in the text only format found online, as show in the examples below.

Positive impoliteness in Counter Speech Examples

Eg.1
P1: Look I'm not racist, but this ferguson bullshit is ridiculous. Just because a black man got killed all the black people are rioting
P2: They think as long as they don't say nigger they're not racist smh

Eg.2
P1: Call me racist if you want. People like Obama, no matter what their color, are the reason for the word nigger to begin with. WORTHLESS POS!!
P2: I always marvel at people who have no problem saying the n word getting shy about spelling out "piece of shit."

Eg.3
P1: I can proudly say as an american I fucking hate our nigger president Obama stupid porch monkey
P2: and oddly enough, capitalized "Obama" but not "america."

Eg.4
P1: Statist/leftist Black Americans: Face it, Obama is a fucking nigger...i.e., a traitor and an embarrassment to your race.
P2: I am trying so hard to wrap my mind around how ignorant this is...

Eg.5
P1: Marriage is a relationship between a man and a woman. I don't think it is the role of the state to define what marriage is.
P2: the. MOST hypocritical thing I have ever heard in my life. Omg I can't stop laughing

Eg.6
P1: Fuck you N ur Obama care u fucking nigger I'll Never do it
P2: Racists and puncuation are never found together.
P3: The sad thing is, stupid can often be controlled with diet & exercise, but so few do.

Fig. 47

Each of these examples shows a counter speaker responding to hate speech in a way that ignores and disassociates the counter speaker from the hate speaker, linguistically excluding them from the discussion that they had started. In all but one example (Eg.1) the hate speaker is not referred to at all, only an analysis in abstract of the messaging is given, and in the one example where they might be referred to, the non-specific "they" pronoun is used to either refer to the subject at a distance or in general with others of their ilk.

Starting with example 1, the repetition of the impersonal pronoun “they” put the counter speaker’s analysis of the hate speaker in abstract and distanced terms, excluding them from the activity of discussing the validity of the idea that to not use the word “nigger” makes someone not racist. The use of the acronym “smh” (shaking my head) work to textually perform the embodied motion of dismissing and condemning this idea and the person presenting it, again without acknowledging them directly. This impoliteness strategy is used, ironically, to dismantle the strategic politeness of the hate speaker to avoid slurs and justify their hate with excuses. The messaging of the counter speech uncovers the tactic of civil language, and the impolite strategy of excluding and ignoring the hate speaker in their response damages their face by designing them as an out-group member compared to the counter speaker and their implied or expected audience.

Examples 2 and 3 once again use positive impoliteness to ignore and show disinterest in the hate speaker, but this time pointing the analytic content of their response at the grammatical inconsistency of the hate speaker and others presumed to be like them. Example 2 identifies the target of their critique as “*people* who…”, abstracting the speaker they are responding to by removing their individuality and consigning them to a group who supposedly share the same poor traits. The only identifying feature the counter speaker affords to the person they are responding to is the inability to properly prioritise their ranking of offensive terms, but again this is something that they suggest is shared by all “people” who can be seen to be part of this out-group. Example 3 uses no pro-nouns or naming language at all in their response, presenting a conversational style that forgoes an expected sentence opening in favour of jumping straight to the ridiculing of the hate speaker’s choice of grammar. The critique presented undermines the hate speaker’s intelligence and the consistency of their message that being “american” is a trait that they believe gives them power and rights which they are proud of. By phrasing these responses in a way which excludes the hate speaker from the discussion, the counter speakers remake the hate speaker not as an equal partner in an exchange, but as a subject to be analysed, critiqued and mocked by the in-group.

Examples 4 and 5 forego any mention of the hate speaker entirely, instead directing their critique and the critique of anyone in the shared in-group audience entirely at their analysis of their linguistic conduct. Both examples use hyperbolic descriptions of the content

of the hate speakers' messages to paint them as "ignorant" (suggesting they are ill informed to the point of incomprehension) and "hypocritical" (explicating the dissonance in internal logic presented in the hate speaker's utterance). This tactic is also shared with example 2, in which the counter speaker critiques the out-group that the hate speaker is associated with but does not address the hate speaker directly. All three of these examples also display an instance of negative politeness in which the targets negative face is attacked by using the personalising pronoun 'I' to "explicitly associate the other with a negative aspect" (Culpepper, 1996: 358) like being ignorant (Eg.4), hypocritical (Eg.5) or showing inconsistent priorities (Eg.2).

Example 6 makes use of these tactics as well, ignoring the hate speaker, disassociating them from the conversation, being unsympathetic, as well as using "derogatory nominations" (Culpepper, 1996: 358) in labelling the counter speaker as suffering from or being "stupid" without addressing them directly through name or pro-noun. This examples also displays the mock politeness strategy discussed before, in presenting an insincere solution to what they are suggesting is causing the hateful speech. In each of these examples the hate speaker's agency and personhood is reduced through this face attack. Providing an alternative narrative or an argument against their hateful rhetoric would provide threat or damage to the speakers face, but by increasing impoliteness additional damage is performed, further undermining them as a source of legitimate knowledge in front of their assumed audience. This reduction in personhood and agency also serves as a counter balancing force to the dehumanising language used by hate speakers against their chosen marginalised target group.

## Impolite Beliefs

### Impolite Beliefs in Counter Speech Examples

Eg.1
P1:Statist/leftist Black Americans: Face it, Obama is a fucking nigger…i.e., a traitor and an embarrassment to your race
P2: Please, leave the country. You're embarrassing the rest of us.

Eg.2
P1: Call me racist if you want. People like Obama, no matter what their color, are the reason for the word nigger to begin with. WORTHLESS POS!!
P2: And you are the reason people call others – uneducated, backward, white trash.

Eg.3
P1: Just saw a gay commercial, #Yuck #Disgusting #Eww #Gross #Nasty #Sick #HowManyHashtagsCanIComeUpWithToDescribeHowAbhorrentHomosexualityIs
P2: lmao, the reason the average human IQ is only 90 is because there are too many people like you beneath that.

Eg.4
P1: rabidly racist blacks will not rest until there's a lynching..fuck them..stand up..
P2: another dumbass racist, who doesn't understand the definition of racism (or lynching). #dumbasfuck
P1: oh and you do of course
P2: Yeah, I do. Cuz I'm not a knuckle-dragging right winger. Let me guess - Fox News viewer?

Eg.5
P1: I'm not a racist but god damn Lil Wayne is a nigger
P2: My condolences to your mother.

Eg.6
P1: mlk didnt die for you nigs to be ignorant im not racist but people like you make me want to be
P1: nigga nigga nigga im going to say it all i want dont want white people to say it then all you shouldnt its not even offensive
P2: This bitch look like a gayer Justin Bieber talking hard.

Fig. 48

This final strategy is derived from an inversion of Leech's politeness model (1983), which Culpeper explains as being focused on linguistic content as opposed to Brown and

Levinson's focus on linguistic form. Culpeper (1996) explains that this strategy for impoliteness "minimize[s] the expression of polite beliefs and maximize[s] the expression of impolite beliefs" (p.358) about the person as a means to damage the targets positive face. It is suggested that face is not simply assigned to properties of the self, but that face exists in layers or "concentric circles" (Liu, 1986) of identity around the self, becoming less face laden as they emanate. Some of these identities can include a person's job, family, nationality, competence, relationships, psychology, intellect etc. By using explicit on-record threats to the multiple layers of face one may have, counter speakers may combine this with more form focused impoliteness strategies to further undermine a hate speaker. In the examples shown above, each counter speaker is avoiding talking about the hate speaker's racism, the thing that caused the offence and encouraged their response, and are instead providing additional impolite beliefs about the hate speaker to heap further face damage upon their target.

Example 1 shows the counter speaker attacking the hate speaker's social role as a member of their nation. This is in direct response to the hateful rhetoric that Obama's status (given by the hate speaker) as a "nigger" equates him with being a traitor to his country and an embarrassment to his race. It is this invocation of patriotism and national alliance which the counter speaker is responding to suggest that the hate speaker should "leave the country" because it is in fact, they who are an embarrassment and not worthy of membership. As noted, before, different layers of group membership or identity will have more or less face attached to them. Because the hate speaker themselves has brough up the idea of national loyalty in labelling the president a traitor, the counter speaker may have presumed this group membership to be something imbued with meaning and importance to the hate speaker. By suggesting they should be ejected from that nation group they are undermining the implied status that membership might provide in the eyes of onlookers and directly attacks the face associated with it by the hate speaker.

Examples 2 and 3 both purport impolite beliefs about the intelligence of their respective hate speakers, with example 2 providing additional judgement on their hate speakers social class and value. Again, these counter speakers both avoid confronting the racism and homophobia presented by the hate speakers, instead providing impolite beliefs about other aspects of their opponents, specifically their lack of intelligence. Example 2

mimics a portion of the original hate tweet to frame an attack on their supposed lack of education, as well as their being “backward” and “white trash”, a racially charged invocation of their social status. Example 3 responds with an explanation that the hate speaker’s lack of intelligence places them in a group which effects the average IQ of the human race, again undermining their argument and attacking their face in terms of intellect and competence to make statements of any legitimacy. While there is no overt indicator in the rhetoric of either hate speaker that perceived intelligence is important to them, any damage to the person’s perceived intelligence will have an effect on the legitimacy with which their ideas and statements are read by others.

The counter speaker in Example 4 once again casts impolite beliefs upon the hate speaker’s intelligence, providing specific examples of their lapse in knowledge by accusing them of not understanding certain words used in their rhetoric. When the counter speaker receives a response that does not concede to their points, they expand upon their accusation of low intelligence by disparagingly supposing the hate speakers’ political alliances. The counter speaker does this through the phrasing “knuckle-dragging right winger” and guessing that they are a watcher of Fox News, a right-wing news organisation. By attaching “Fox News” as an information source to the hate speaker, the counter speaker is identifying them as misinformed and biased to those who would hold those suspicions of Fox News and by attaching “knuckle-dragging” to their description of right-wing political beliefs they make explicit their judgement of that ideology. So then, these impolite beliefs expand on the attack on intelligence to include an attack on their critical thinking in terms of news source and on their political identity. While the hate speaker may be unashamed or even proud of their right-wing political beliefs, their equating to “knuckle-dragging” and ignorance will provide face damage to them and enhanced undermining to others watching and agreeing with the counter speaker.

Example 5 takes aim instead at the familial role of the hate speaker, characterising the hate speaker not as unintelligent but as a disappointment or inconvenience to their mother. The counter speaker here makes no mention of the content of the hateful rhetoric, and makes no explicit references to the hate speaker, instead providing only an implication that the hate speaker must be bad or wrong through their feelings of condolence to their parent based on

their current conduct. Parent/child relationships are often assumed to be of high importance and the invocation of the mother in particular during insults and ritual putdowns is very common (Abrahams, 1962, Terrion and Ashforth, 2002, Murphy, 2017) and it is used here to perform an impolite belief that moreover than mere judgement from a stranger on the internet, their choice of rhetoric and displayed belief would be shameful and embarrassing to their mother.

The final example shows a counter speaker again making a deliberate attempt at damaging the hate speakers face, but this time employs misogynistic and homophobic language to denigrate and attack the hate speakers assumed masculinity. Interestingly and seemingly without self-awareness, the counter speaker in this instance uses hateful rhetoric themselves in an attempt to directly damage the face of the hate speaker, describing them as a “bitch” and looking like a “gayer Justin Bieber” to discount the idea that they can “talk[ing] hard” about the right of white people to say the word “nigga”. While the use of homophobia and misogynist language may in fact undercut the counter speakers’ message in the eyes of some onlookers, the counter speaker appears to be using that language to paint the hate speaker as weak, unintimidating and because of that unworthy of any legitimacy that their statement may try to claim. The counter speaker is presuming heteronormativity, masculinity and strength to be important components of the hate speakers positive face because of the brash and unapologetic manner with which they wield a word, which owing to their offered explanation, they must be aware carries controversy. The counter speaker then has decided to forego a discussion of their logical or ethical legitimacy to the use of that word and attack another aspect of their face to cause damage and as a means to undermine them in the eyes of others who might also give worth to those attributes.

## Summary

This chapter explored different methods of linguistic impoliteness, deployed by counter speakers to add power to the rhetoric used to achieve their communicative aims. Particularly, this chapter has grappled with the use of impolite language by those seeking to protect targets of hate. By using Culpeper’s (1996, 2003, 2005) inversion of Brown and

Levinson's seminal work to devise a taxonomy of impoliteness strategies, this chapter has shown examples of these strategies being deployed in the wild to actively and deliberately deliver damage to the face of those hate speakers, the face of anyone who would think to agree with them, and to undermine their ideas to those undecided who make up the presumed audience of online interaction. By engaging in mock politeness, positively impolite ignoring and impolite beliefs, the counter speakers exemplified here show tactics which seek to enhance face damage to their opponent as a means to undermine their message and provide protection for those at risk.

# CHAPTER 10: Conclusions

## INTRODUCTION

This final chapter draws together the findings from the analysis presented in the previous six chapters, discussing the key themes that occur within and between hate and counter speech with regard to the original research questions set out in the opening of the thesis. In performing a computer mediated discourse analysis of naturally occurring online hate and counter speech, this thesis has attempted to generate a unique and novel contribution to the existing field of study, deepening the understanding of how hate and counter speech are created, justified, read and combatted online. This chapter will address the key research questions that drove this research, pulling out key themes and findings discovered throughout the analysis. Finally, this chapter will draw conclusions from the study as a whole, before suggesting tentative policy applications, further research, and reflecting on the limitations of the data and analysis.

## ANSWERING THE RESEARCH QUESTIONS

1. How are the communicative aims of hate speech achieved online, in spite of technological interception and social stigma?
2. How are the communicative aims of counter speech achieved online?
3. How is identity created, demonstrated and weaponised to achieve the communicative aims of hate speech and counter speech?
4. In what way are (a) politeness and (b) impoliteness manufactured and used as a tool for the production of hate and counter speech?
5. How is the intertextual gap between "generic" hate/counter speech and "naturally occurring" hate/counter speech manipulated?

To properly answer questions 1 and 2 of this research, a working definition of each of these types of language had to be discerned from which to glean their communicative aims. These questions are not concerned necessarily with the semantic tropes or specific language

that typifies these discourses, but what they aim to communicate and achieve through their language.

In synthesising the many concurrent definitions of hate speech discussed in the literature review, this thesis suggests that hate speech was defined by 4 typical features; its identification of a vulnerable group (or individual chosen because of their membership to that group), its negative description of that group, injurious speech aimed at the target, and an incitement for others to confirm and share those injurious sentiments. The communicative aim then is to injure or other an individual or group based on their vulnerability and broadcast that othering rhetoric to convince people of its veracity.

Counter speech, owing to its comparative lack of widespread definitions was more difficult to define in as much detail. However, ground-breaking work by Williams (2021, Richards and Calvert (2000) and Wright et al. (2017) provided a base from which to generate a useful conception of counter speech. While the specifics of the operation of counter speech were not consistent across all definitions (some calling for "counter narratives" (Ernst et al., 2017) and some happy to accept any 'direct response' that requires no law or institution (Wright et al., 2017)) the key communicative functions of counter speech showed themselves to be an attempt to undermine hate speech, invalidate the positions put forward, and reduce harm to the target.

As noted above, each of these discourses are seen to be typified by a different focus in the ways they achieve their aims. While some of the hate speech studied was explicit and brazen, it was almost always accompanied by some sort of mitigating hedge which aimed to justify their speech, no matter how offensive. The presence of these hedges illuminated the fact that there was indeed a social stigma to broadcasting hate speech, even if a technological one was not available. And it is those hedges which sought to avoid the repercussions of that social stigma wherever possible. The preferred method found in the data for avoiding that social stigma appeared to be authorisation. In each of the three analytic lenses, hate speakers were found to employ different rhetorical moves which positioned themselves as imbued with authorial power. Rather than shrinking or softening their rhetoric, hate speakers were often found performed as if they were presumed correct, justified and empowered by

consensus, and acted with the intention that this performance would overcome any issue raised by the hateful speech that came along side.

Counter speech, however, achieved its aims in a different way. Rather than being encumbered by social pressure or algorithmic policing, the only presumable barriers to involvement were the fear of repercussion or a lack of ability. Both of these possible barriers may influence the techniques found to be recurrent in the data, those typically involving some sort of humorous element (joke making, absurdity, mimicry and sarcasm), or the repurposing of information already supplied by the hate speaker (mimicry, reductio ad absurdum, oppositional identity). As discussed, humour has been shown to have a mitigating effect when discussing taboo or tense subjects, and also a community building effect (Terrion and Ashforth, 2002) in creating a collaborative "play frame" (Coates, 2007), in which collaborative intent is understood and a target is agreed upon. Both of these features could go some way to reducing the fear of reprisal a would-be counter speaker might feel. In instigating a humorous tone, their retort may be assumed to be taken less aggressively than otherwise and may allow for others to join and group safety to be increased. Alternatively, in the application of mimicry, reductio ad absurdum and oppositional identity, the counter speaker is able to create counter narratives and responses with very little of their own knowledge required. By recreating the tone and structure (Richardson et al., 2014) of a hateful utterance, or extending a hateful premise to its logical conclusions (Novaes, 2016) with the aim of amplifying the issues inherent in them, counter speakers are able to design responses to the hateful content as well as the speaker, with immediate effect.

Question 3 sought to understand identity work in hate and counter speech, and this analytic frame identified 3 distinct forms of identity work within the data; declarative (Silverman, 1998; Goffman, 1967), oppositional (Benhabib, 1996; Connolly, 1991) and weaponised (van Leeuwen, 2007; Lopez, 2014; Alberston, 2014).

While all three forms of identity work were seen amongst hate and counter speakers, each genre tended towards their own preference. Invoking Sacks' "inference rich categories" (1995), hate speakers were comparatively much more likely to make declarative statements about their identity, stating who they were, where they were from and what about their

character gave them the authority (van Leeuwen, 2007) to proclaim their ideas and opinions. While these statement identities were seen often, hate speakers also frequently engaged in more subtle oppositional identified. Occasionally as a means to bolster their declared identity, hate speakers were found to place themselves in a privileged standard category by way of distinguishing the "other" category as lesser. By categorising black people, gay people and immigrants as "them" and worthy of disparagement, they reinforce their own identity as standard, local and morally beyond reproach. However, for all the work done by this identity management, much of it seemed to pale in comparison, and be undone by, the identity performed through racist rhetoric. This may be a side effect of the weapon focus (Loftus, Loftus and Messo, 1987) from being highlighted by a sentinel account outing hate speakers, but the respondents found replying were far more convinced of their identity as a racist or homophobe than they were by the prestige of being anything else.

Counter speakers on the other hand, despite the occasional use of declaration, were found to often forego an explicit identity creation, in favour of defining themselves by what they oppose. They again were found aligning with van Leeuwen's (2007) concepts of legitimation, this time providing moral evaluation through loaded descriptors such as "pathetic" or "racist". The counter speakers' identities almost seem incidental, in the face of delegitimising the hate speaker and re-constructing them as less than they portrayed. This also leads into the counter speakers frequent use of weaponised identities, where identities are rejected, remade and deployed to damage their face and the power of their speech. By providing inference rich category markers like "child" or "low class", counter speakers design an identity for the hate speaker and deploy it with the intention that this new identity does the work to undermine their rhetoric.

Question 4 sought to understand the seemingly contradictory discovery that hateful speech is often presented in a decidedly polite way, while counter speech was often very impolite. These questions intended to uncover what communicative function politeness or impoliteness had in conveying their types of speech effectively.

Hateful speech, despite occasionally containing slurs and rabid anger, is frequently presented as calm and courteous, almost apologetic in its deployment. This analysis identified

three different politeness techniques used by hate speakers, and found that politeness in this case is not always produced in a two way interaction, but often is used to address the third party in online discourse – the presumed audience. Firstly, Planchenault's (2010) discussion of 'solidarity markers' as positive politeness proved particularly relevant as it was used frequently in the data to denote in-group togetherness. This feature was also reminiscent of Mehlman and Snyder's (1985) consensus raising, where hate speakers in the data invoked an assumed "we" when proclaiming their judgements on their chosen targets. While this othering by opposition increases face damage to the target, an impolite act, it simultaneously performs positive politeness to those around the interaction, who may feel welcomed into the dominant in-group that has been generated. By extending the individual motivations of their speech out to the "us" in-group the hate speaker also disperses the responsibility for the damage, and also legitimises it as part of the social contract of consensus (Rousseau, 1994).

The next form of politeness identified in the data was going "off record", a technique designed for softening the impact of hate speech against the target. By increasing ambiguity through the use of irony, metaphor and rhetorical questions, hate speakers provide an out for their target to be allowed to interpret their hate as a misunderstanding. One can choose to read a racist dog whistle as an honest discussion about rates of black on black crime, or see homophobia as misguided concern for children if that reading will protect and mitigate face damage.

Lastly, redressive action was shown to be another preferred method of politeness, used by hate speakers to mitigate their speech. Through redressive action, hate speakers provide an attempt to partially satisfy the negative face wants (Brown and Levinson, 1978) of their target, minimising the damage through apology, self-effacement or impersonalisation. Hate speech in the data was often found to be front loaded with an apology, placed there to reduce damage and provide a small buffer before the hateful content was delivered. By stating some faint praise about a marginalised group before indicating that they are not equal and accepted as those in the privileged standard category, the hate speaker's politeness softens the blow somewhat. This allows the contrary statement to be taken as an invalidation of the hateful label that would be otherwise applied without it.

Counter speech, ironically, is often very impolite, containing the attacks and stigmatizing labelling that were mentioned earlier. The concept of impoliteness that drove this analysis was informed primarily by the work of Culpeper (2005), whose inversion of Brown and Levinson's (1978) politeness framework provided a working definition of impoliteness as not merely accidental incivility, but active attempts at face damage.

Sarcasm and mock politeness were the first concepts that made themselves known in the data. The analysis found frequent instances where counter speakers went out of their way to insert obviously sarcastic polite responses that were plainly insincere. Because of the lack of prosody in online text, cases of mock politeness were made very explicit by oddly formal phrasing, rhetorical questions and in some cases innovative use of punctuation to suggest tone. Many of the examples shown in the analysis have non-sarcastic counterparts in the politeness analysis, but these impolite instances are shown to be insincere through their incongruous context and odd form.

Next, positive impoliteness was found recurrent in the data, a method whereby the positive face wants of the target, to be treated as a friend or in-group member, are damaged through snubbing, exclusory language, disassociation and disinterest. Because it is hard to demonstrate ignoring online, owing to the need to actively engage with a person to have any impact, snubs in this context were mostly performed by replying to a hateful utterance in a way that suggests the respondent isn't addressing the speaker at all, but is talking to the in-group or themselves. In framing their response this way, many counter speakers engaged in insults and harsh criticism in a style that suggested they were talking *about* the hate speaker, rather than to them.

Lastly, the impolite beliefs strategy, which seeks to identify more "face laden" identity characteristics of a person and institute poor beliefs about them, was noted within the data. This technique uses bald on record face attacks, but is designed to focus on parts of the identity which may be more hurtful to damage, such as their job, their family, their intellect etc. This technique takes as its central tenant an altered form of Goffman's face concept, suggesting that face works in concentric circles, with each one closer to the centre being more

emotionally attached and vulnerable to wounding damage. Impolite counter speakers were found ignoring the content of the hate speech and instead jabbing at the national pride, family ties or intelligence of the hate speaker.

Question 5 examined the difference between "generic" hate and counter speech (that is, speech which explicitly falls within the boundaries set out by its genre classification), and the 'naturally occurring' hate and counter speech found in the data. Obviously, some or all of the data collected could engage in hate speech that does align perfectly with the genre conventions they are defined by, but natural human discourse rarely follow quite so succinctly. These questions sought to understand what people online did to achieve the communicative aims of their genre of speech outside of the standard tropes found in the definitions.

Many of the individual techniques that were found to manipulate the intertextual gap between generic speech and natural speech have been touched upon in the discussions of identity and (im)politeness. Each of these have shown ways in which speech has been mitigated, altered, or re-oriented to achieve some particular communicative aim within the hate and counter speech genres.

Within the hate speech analysis, the core techniques used to manipulate the genre were legitimation, excuse making and projection. As discussed before, legitimation was found repeatedly in the data, with hate speakers employing different tactics to legitimise their speech, give themselves authority and move their utterance beyond question. By invoking nationality, religion or a genuine critique, hate speakers attempted to legitimise their hate speech through those associations and avoid questioning.

Excuse making performed a few different functions in manipulating speech. Firstly, it attempts to raise consensus, suggesting that the poor performance they are excusing is the same outcome that any other actor would experience in the same situation. By raising consensus, hate speakers reduce their responsibility and reduce the immorality of their actions by proposing its normalcy. An alternative function of excuse making is distinctiveness raising, which contrastingly seeks to separate this instance of poor performance as distinct

from pervious or future attempts at performance, reducing its generalisability. The separation of responsibility for this poor performance also redirects any criticism from the person directly and onto the act itself. Lastly, consistency lowering works to indicate a lack of consistent performance across time and activity, suggesting that what occurred this time and all others is due to outside influence and external factors. All of these were shown to be deployed by hate speakers, in an attempt to manipulate their speech away from the prescribed generic trappings of hate speech, but still achieve the aims which retaining less of the responsibility.

Lastly projection was found to be a mitigating tactic in the deployment of hate speech. In this tactic, illocutionary directive speech is used to relocate the moral responsibility for the production of hate speech away from the speaker and onto an awaiting audience. By phrasing hateful sentiments as questions rather than statements, hate speakers work to place the responsibility for confirming or denying those sentiments away from themselves. By designing questions, hate speakers preload their ideas with assumptions. Namely the assumption that there is an audience member listening, and that they possess the capacity to engage in the discussion they are presenting. Additionally, in creating Idealized Cognitive Models (Hernández and Mendoza, 2002) through the presentation of a small snapshot of information and following it up with a binary choice option, hearers of this projection are often left with little option and nowhere to divert discussion away from the hateful but obvious choice.

Each of these tactics are suggested as methods for using language to manipulate the intertextual gap between what is being said and the aims of hate speech they intend to achieve. By using intertextual queues to this and other genres of speech, the lines are blurred enough to bring doubt upon those who would condemn that form of speech, while simultaneously alerting those who would agree.

The three core aspects of the genre analysis of counter speech are comedy, mimicry and reductio ad absurdum. Each of these was found to help achieve the generic aims of counter speech, as suggested by their definition, but outside of the classical proscribed techniques such as calls for information (Williams, 2019).

In comparison with hate speech, a great deal of the linguistic techniques identified in the counter speech analysis were based around the inclusion of joke and comedy into their response. There was a high frequency of humour, irony, sarcasm, mimicking and snubbing found within the counter speech responses analysed, which may suggest that where hate speakers are making efforts to identify themselves as serious sources of insight, counter speakers may be making attempts to defile that self-important identity, or as Terrion and Ashforth (2002) and Coates (2007) might suggest, are seeking to engender temporary group solidarity through joke and ritual put downs.

Rather than providing alternative narratives or countering information to tackle bad claims head on (as prescribed for more high-quality counter speech), much of the counter speech investigated in this study used comedic methods to attack and delegitimise the identity performed by the hate speaker. While this may be classed as antagonistic, this kind of delegitimising speech can produce some relevant impacts and may have more pragmatic motivations.

Beginning with the effects that may be caused through the face damaging responses, primarily these attacks appear to aim at delegitimising the hate speaker. Without addressing the specific content which identified the speaker as hateful, many counter speakers are able to generate convincing arguments, both moral and mischievous, as to why the statements made by the hate speaker should be ignored. What is more, in delegitimising the hate speaker, their power and authority are reduced and thus they may become less intimidating an opponent to other would-be counter speakers. When onlookers become aware that an authoritative sounding hate speaker is susceptible to face attacks, and that those can be made without a wealth of opposing knowledge, they may themselves feel emboldened to become involved in counter speech. Williams (2019, 2021) notes repeatedly that as much as high quality counter narrative speech is preferred, the effect of timely opposition from a high volume of speakers must not be understated.

This leads on to the pragmatic motivations of engaging in this form of counter speech. As noted, counter speakers are created in the moment, borne from their opposition to a

hateful utterance. They do not prepare for battle with their opponents but are drawn in spontaneously when issues arise. Undoubtedly, some counter speakers are more seasoned and may have collected useful countering information over their encounters, but many or most users who stumble across abhorrent hate speech will be untrained and improvising. It is for this reason that alternative tactics, like absurdist asides, comedic roasts and mimicry may appear so attractive to new and would be counter speakers. In deploying these techniques which require far less prepared knowledge and in some cases are entirely created by repurposing the hate speakers' content, counter speakers are able to engage hateful material as it occurs far more immediately and in greater numbers.

This is not, however, to suggest to ignore the fact that counter speakers may themselves be driven by the anger of injustice, and engage in face damaging attacks for the sake of wounding the hate speaker. While counter speech as a discursive tool does contain the potential to be a reintegrating force (Braithwaite, 1989), stemming hate speech and convincing speakers to the error of their ways, it can also be a plainly stigmatising tool. Much of the discussion in this thesis around humour being a mitigating feature in providing a more reintegrative form of shaming, however, it may be worth noting the possible effect of stigmatising a 'confirmed' visible hate speaker for the echo of reintegrative shame it might provide to onlookers who might have convinced to hate. In this way, one might view the more blunt and traumatic action of face attacking a hate speaker as cutting off the head of the snake in the hope that it might immobilise and reform the body.

The next technique identified is mimicry. An interesting and frequent phenomenon found in this data was the use of structural and linguistic mirroring by counter speakers. Again, there may be a pragmatic or logistic motivation behind this, this may be a quick and convenient way to design a response that will fit within the boundaries of the platform they are on since one just like it has already been deployed. This may also be a way to design a response that does not require an entirely novel pre-knowledge on the subject at hand. However, what this does do beyond those things, is create a framework in which the similarity of the response amplifies those few changes which are different, narrowing the focus of the hate speaker and the audience around them to be able to see specifically what is wrong with their hate speech. By changing one word or another, immediate comparisons can be made, and when those

changes are to something absurd or extreme, the logic behind the surrounding argument is illuminated and exposed for any flaws it may have.

Mimicry as a linguistic technique illuminates the co-constructed nature of these genres of speech, particularly counter speech whose who existence and purpose relies on the that which it opposes. Mimicry most explicitly is a linguistic which necessitates a 'collaborative' procedure, where the structure and language of a response is generated through the interplay between interactants. Co-construction is built into the fabric of interaction, collaboratively generating knowledge (Dubovi and Tabak, 2020) and understanding, and creating and reifying identity (von Wallpach et al., 2017), and in the online digital space that notion appears amplified (Calcagni et al., 2019), where identity and meaning are made and recreated constantly, hampered by anonymity and empowered by technology.

Finally, reductio absurdum takes a similar but distinct stance as mimicry, using the content provided by the hate speaker to provide an alternative version which pushes their assumptions to their extreme logical conclusions as a means to highlight their flaws. This form of counter speech once again avoids the generic trappings of prescribed counter speech tactics, but allows the hate speaker and the audience to unpack for themselves the issues with the hateful rhetoric by pushing it to an extreme that the hate speaker could never rightfully defend or justify.

Again, each of these linguistic techniques were identified as ways that users were able to achieve the communicative aims of counter speech without directly engaging in the generic tropes that are prescribed by its definition. Because counter speech is less tightly defined, and less policed in general, the manipulation here is not used to avoid social or architectural barriers to achieving their aims, but the methodological and logistical ones that come with trying to engage in a highly contentious but highly impromptu form of online discourse. Each of these techniques aids users in avoiding and overcoming those barriers in their attempt to be a "hate speech first responder" (Williams, 2021).

In undertaking this research, this thesis has worked to contribute to the knowledge base and understanding of the linguistic and rhetorical structure for both hate and counter speech. By

engaging in a fine-grained computer-mediated discourse analysis of online hate speech and its responses, this thesis has endeavoured to discover the key linguistic strategies deployed by users in the wild in the production of hate and the strategies most readily available to those who wish to combat it. While not exhaustive, this novel knowledge may contribute towards a toolkit for the in-situ identification of hate speech tactics, as well as the building blocks for on-the-fly responses constructed from the content and contextual information offered by hate speakers. It is not the intent of this thesis to suggest that counter speech is the ultimate and most effective weapon in the battle against online hate speech, but to acknowledge that with an ever-expanding online space for the production and policing of hate speech, an informed and responsibilized citizenry is more necessary than ever. This thesis hopes to provide a thorough and useful contribution towards the understanding of hate speech and expanding the ability of counter speakers online.

## LIMITATIONS

This thesis, as with any piece of academic work, has come up against the limitations necessitated by its theoretical and methodological choices.

In focusing this investigation on a qualitative discourse analysis, it has come up against limits which will affect the external validity and reliability. This study was conducted through the interpretation of one researcher, and this analysis will suffer from the biases of that research, which are sure to be numerous when dealing with content of this nature. While the analyst's reading of the data may view speech as hateful, and interactants within the data may identify speech as hateful, the producer of that speech themselves may not. Perspective differences such as these will always have an impact in qualitative research and may colour the interpretation of what linguistic techniques are being used and what ends they are trying to achieve with those techniques. Similarly, what is coded and analysed as counter speech may seem apparent, but could well be interpreted in many instances as attacking, harassing, or even hateful speech in itself. Additionally, although there was a fairly large collection of data that was engaged with, in the grand scheme of online big data studies, it pales in comparison. By not expanding the scope of this study to include comparative data sets from

different times, different platforms and different sampling methods, this research is no doubt restricted in its ability to be generalised into other settings.

Another limitation encounter by this study during the collection and sampling phase was the restriction in multimodal representation of the data. A methodological driver of this research was the EM/CA focus on in situ analysis of methods in interaction, and that was well suited by the lack of demographic information which arose from having no profiles, bio's or pictures accompanying the data as it was presented. However, because the data was presented as pure text occupying a spread sheet, the data was missing additional forms of communication and interaction such as emoji, picture, Gif and video. Online interaction is becoming more textured with additional media being seamlessly integrated into blogs and posts, and this research was unfortunately unable to incorporate those rich data sources into its analysis.

Finally, the general speed with which social media and online data develops means this kind of research is always necessarily catching up with the current landscape. In the time since this research began, there have been two U.S. presidents, two UK general elections, an American insurrection, a global pandemic, the accelerated prominence of the Black Lives Matter movement and a disturbing uptick in global anti-Asian hate crime, to name a mere handful of the events have the influenced online hate and counter speech. All of this and more has been filtered through digital platforms, changing not only the tone and topic of online discussion, but forcing architectural change to the platforms those discussions occur on. However, this thesis hopes to provide a unique snapshot of the context of its undertaking with the aim of providing some modicum of a novel contribution, from which further research can be developed.

## SUGGESTIONS FOR FURTHER SCHOLARSHIP

As noted, there are a myriad of limitations which could be addressed in future research which would provide texture and triangulation to what was discussed here. This thesis itself is a continuation of two previous masters' dissertations, each of which laid some

theoretical groundwork from which this was designed. The intention of this thesis was to provide some of the qualitative detail which seemed to be lacking in hate and counter speech research and interrogate some of the assumptions taken for granted around the ways these discourses work in everyday digital life.

In terms of further scholarship based around this research specifically, it may be advantageous to engage in a replicated or semi-replicated study, engaging with the data in an attempt to discover different recurrent linguistic techniques, as informed by different tools within the CMDA toolbox.

Useful future research might also consider transplanting these qualitative concepts back over to a quantitative analysis, employing machine learning or AI to identify these linguistic techniques on a grander and more contemporaneous scale. There is a strong and necessary partnership between the qualitative and quantitative aspects of this area of research and it is a hope of this thesis to sit as a steppingstone towards one such collaboration.

Additionally, the findings of this research could be used to inform a study identifying the effectiveness of different kinds of counter speech techniques on different forms of hate speech. In deploying techniques like mimicry or absurdity in response to different kinds of naturally occurring hate speech, a further study could track the efficacy of those interventions to identify whether it leads online hate speakers towards desistence, or if certain techniques exacerbate and inflame certain actors.

## POLICY RECOMMENDATIONS

While the interpretive nature and specificity of this thesis may preclude it from offering any official policy recommendations, it would put forward a suggestion for co-operation with third sector institutions and charities such as Stop Hate UK in designing educational literature for identifying hate speech, conducting counter speech and encouraging engagement more widely. In better understanding the techniques commonly

used by counter speakers in the wild, charity organisations and anti-hate institutions could work towards creating 'generic' templates or examples of some of these forms of counter speech, or provide training in how to produce authoritative identity work in their counter speech to bolster its effectiveness.

Additionally, in breaking down some of the techniques used to enhance the power of hate speech, educational materials could be produced which help to catalogue these techniques for easy identification in the wild. In understanding how hate speakers convince an audience with linguistic flourishes and identity work, social media users may be better equipped to reject those ideas and see past the hedges to the hateful content they are trying to justify.

If further research could confirm the efficacy of these counter speech tools, they need not be restricted to institutional use, but may be useful as anti-hate education for all online citizens. As has been discussed throughout this thesis, one of the touted advantages of counter speech is its ability to be performed by anyone to maintain their communities, combat intolerance and reduce harm wherever possible. Hopefully the insights described in this research can provide some aid in forwarding that goal of encouraging a community of counter speakers to use their voice as frequently and effectively as they can.

# Bibliography

Abrahams, R. 1962. "Playing the Dozens.." *The Journal of American Folklore* 75(297), pp. 209-218. doi: 10.2307/537723.

Adam, B. 2003. When Time is Money: Contexted Rationalities of Time in the Theory and Practice of Work. *Theoria: A Journal of Social and Political Theory* 50(102), pp. 94-125. doi: 10.3167/004058103782267403.

*Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems 2003* Strasbourg: Council of Europe.

Albertson, B. 2014. Dog-Whistle Politics: Multivocal Communication and Religious Appeals. *Political Behavior* 37(1), pp. 3-26. doi: 10.1007/s11109-013-9265-x.

Angouri, J. and Tseliga, T. 2010. “You Have No Idea What You are Talking About!”: From e-disagreement to e-impoliteness in two online fora. *Journal of Politeness Research. Language, Behaviour, Culture* 6(1). doi: 10.1515/jplr.2010.004.

*Anti-terrorism, Crime and Security Act 2003* London. c.24.

Archakis, A. et al. 2018. “I’m not racist but I expect linguistic assimilation”: The concealing power of humor in an anti-racist campaign. *Discourse, Context & Media* 23, pp. 53-61. doi: 10.1016/j.dcm.2017.03.005.

Atkinson, P. 2013. Ethnography and Craft Knowledge. *Qualitative Sociology Review* IX(2), pp. 56-63.

Attardo, S. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics* 32(6), pp. 793-826. doi: 10.1016/s0378-2166(99)00070-3.

Ayo, F. et al. 2020. Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. *International Journal of Intelligent Computing and Cybernetics* 13(4)

Baron, N. 2008. *Always On: Language in an Online and Mobile World*. Oxford: Oxford University Press.

Bauman, Z. 2007. *Liquid Times: Living in an Age of Uncertainty*. Cambridge: Polity Press.

Baym, N. 2010. *Personal Connections in the Digital Age*. Cambridge: Polity Press.

Beccaria, C. 1764. *On Crimes and Punishments*. Indianapolis: Hackett Publishing Company.

Beetham, D. 1991. *The Legitimation of Power*. Hampshire: Palgrave Macmillan.

Benhabib, S. 1996. *Democracy and difference: Contesting the boundaries of the political*. Princeton NJ: Princeton University Press.

Benwell, B. and Stokoe, E. 2006. *Discourse and Identity*. Edinburgh: Edinburgh University Press.

Billig, M. 1990. Stacking the Cards of Ideology: The History of the Sun Souvenir Royal Album. *Discourse & Society* 1(1), pp. 17-37. doi: 10.1177/0957926590001001002.

Billig, M. 2001. Humour and Embarrassment: Limits of 'Nice-Guy' Theories of Social Life. *Theory, Culture & Society* 18(5), pp. 23-43. doi: 10.1177/02632760122051959.

Billig, M. 2001b. Humour and Hatred: The Racist Jokes of the Ku Klux Klan. *Discourse & Society* 12(3), pp. 267-289. doi: 10.1177/0957926501012003001.

Billig, M. 2005. *Laughter and Ridicule: Towards a Social Critique of Humour*. 1st ed. London: Sage.

Bishop, L. and Gray, D. 2017. Ethical Challenges of Publishing and Sharing Social Media Research Data. In: Woodfield, K. ed. *The Ethics of Online Research: Volume 2*. Bingley: Emerald Publishing Limited

Blumer, H. 1969. *Symbolic Interactionism: Perspective and Method*. 1st ed. Berkeley: University of California Press.

Bonilla-Silva, E. and Forman, T. 2000. "I Am Not a Racist But...": Mapping White College Students' Racial Ideology in the USA. *Discourse & Society* 11(1), pp. 50-85. doi: 10.1177/0957926500011001003.

Bourdieu, P. 1986. The forms of capital. In: Richardson, J. ed. *Handbook of Theory and Research for the Sociology of Education*. London: Greenwood Press

Bousfield, D. 2008. Impoliteness in the struggle for power. In: Bousfield, D. and Locher, M. ed. *Impoliteness in Language: Studies on its Interplay with Power in Theory and Practice*. Berlin: Mouton de Gruyter

Braithwaite, J. 1989. *Crime, shame, and reintegration*. New York: Cambridge University Press.

Briggs, C. and Bauman, R. 1992. Genre, Intertextuality, and Social Power. *Journal of Linguistic Anthropology* 2(2), pp. 131-172. doi: 10.1525/jlin.1992.2.2.131.

Brown, P. and Levinson, S. 1978. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Bucholtz, M. and Hall, K. 2005. Identity and interaction: a sociocultural linguistic approach. *Discourse Studies* 7(4-5), pp. 585-614. doi: 10.1177/1461445605054407.

Burchell, G. et al. 1991. *The Foucault effect*. London: Harvester Wheatsheaf.

Burnap, P. et al. 2014. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining* 4(1). doi: 10.1007/s13278-014-0206-4.

Burnap, P. and Williams, M. 2015. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet* 7(2)

Butler, J. 1997. *Excitable Speech: A Politics of the Performative*. New York: Routledge.

Button, M. and Whittaker, J. 2021. Exploring the voluntary response to cyber-fraud: From vigilantism to responsibilisation. *International Journal of Law, Crime and Justice* 66. doi: 10.1016/j.ijlcj.2021.100482.

Calcagni, F. et al. 2019. Digital co-construction of relational values: understanding the role of social media for sustainability. *Sustainability Science* 14(5), pp. 1309-1321. doi: 10.1007/s11625-019-00672-1.

Cardozo, B. et al. 2003. Mental health, social functioning, and feelings of hatred and revenge of Kosovar Albanians one year after the war in Kosovo. *Journal of Traumatic Stress* 16(4), pp. 351-360. doi: 10.1023/a:1024413918346.

Carlin, A. 2021. Sacks's plenum: the inscription of social orders. In: Smith, R. et al. ed. *On Sacks: Methodology, Materials, and Inspirations*. Oxon: Routledge

Carlson, G. 2016. *I'm Not Racist, I Love Those People: How Trump's Language Reveals His Bigotry*. Augustana Digital Commons.

Carr, P. 2010. The problem with experimental criminology: A response to Sherman's 'Evidence and Liberty'. *Criminology & Criminal Justice* 10(1), pp. 3-10. doi: 10.1177/1748895809352589.

Castells, M. 1996. *The Rise of the Network Society*. Chichester: Wiley-Blackwell.

Chakraborti, N. and Garland, J. 2012. Reconceptualizing hate crime victimization through the lens of vulnerability and 'difference'. *Theoretical Criminology* 16(4), pp. 499-514. doi: 10.1177/1362480612439432.

Chakraborti, N. and Zempi, I. 2012. The veil under attack: Gendered dimensions of Islamophobic victimization. *International Review of Victimology* 18(3), pp. 269-284. doi: 10.1177/0269758012446983.

Chandrasekharan, E. et al. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. In: *The ACM on Human-Computer Interaction*. ACM Digital Library. Available at: https://dl.acm.org/doi/10.1145/3134666 [Accessed: 16 June 2021].

Chiles, A., 2019. *There's a simple way to curb the trolls – end their anonymity | Adrian Chiles*. [online] The Guardian. Available at: https://www.theguardian.com/commentisfree/2019/apr/11/simple-way-to-curb-trolls-end-anonymity-adrian-chiles [Accessed 2 June 2021].

Coates, J. 2007. Talk in a play frame: More on laughter and intimacy. *Journal of Pragmatics* 39(1), pp. 29-49. doi: 10.1016/j.pragma.2006.05.003.

Cohen, L. and Felson, M. 1979. Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review* 44(4), pp. 588-608. doi: 10.2307/2094589.

Committee of Ministers 1997. *Recommendation No. R (97) 20 of the Committee of Ministers to Member States on "Hate Speech"*. Council of Europe.

Community Standards | Facebook. 2020. Available at: https://www.facebook.com/communitystandards/hate_speech [Accessed: 10 June 2020].

Connolly, W. 1991. *Identity/difference. Democratic negotiations of political paradox*. London: Cornell University Press.

Cornish, D. and Clarke, R. 1987. Understanding Crime Displacement: An Application of Rational Choice Theory. *Criminology* 25(4), pp. 933-948. doi: 10.1111/j.1745-9125.1987.tb00826.x.

Cotos, E. et al. 2015. Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes* 19, pp. 52-72. doi: 10.1016/j.jeap.2015.05.004.

*Crime and Disorder Act 1998* UK Public General Acts: The National Archives. c.37.

Crime Survey for England and Wales 2020. *Hate Crime, England and Wales, 2019/20*. London: Home Office.

*Criminal Justice Act 2003* UK Public General Acts: The National Archives. c.44.

Crown Prosecution Service 2020. *Racist and Religious Hate Crime - Prosecution Guide*. London: CPS.

Culpeper, J. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics* 25(3), pp. 349-367. doi: 10.1016/0378-2166(95)00014-3.

Culpeper, J. 2005. Impoliteness and Entertainment in the Television Quiz Show: The Weakest Link. *Journal of Politeness Research. Language, Behaviour, Culture* 1(1), pp. 35-72. doi: 10.1515/jplr.2005.1.1.35.

Davis, T. 1996. *Billy Madison*. Hollywood: Universal Pictures.

Dubovi, I. and Tabak, I. 2020. An empirical analysis of knowledge co-construction in YouTube comments. *Computers & Education* 156. doi: 10.1016/j.compedu.2020.103939.

Duranti, A. 1984. Lauga and Talanoaga: Two Speech Genres in a Samoan Political Event. In: Brenneis, D. and Myers, F. ed. *Dangerous Words: Language and Politics in the Pacific*. New York: New York Universitiy Press

Eckert, P. and McConnell-Ginet, S. 2003. *Language and gender*. Cambridge: Cambridge University Press.

Ede, L. and Lunsford, A. 1984. Audience Addressed/Audience Invoked: The Role of Audience in Composition Theory and Pedagogy. *College Composition and Communication* 35(2), p. 155. doi: 10.2307/358093.

Ernst, J. et al. 2017. Hate Beneath the Counter Speech? A Qualitative Content Analysis of User Comments on YouTube Related to Counter Speech Videos. *Jounral for Deradicalization* 10

Fairclough, N. 1994. Conversationalization of public discourse and the authority of the consumer. In: Keat, R. et al. ed. *The Authority of the Consumer*. London: Routledge

Ford, T. et al. 2008. More Than "Just a Joke": The Prejudice-Releasing Function of Sexist Humor. *Personality and Social Psychology Bulletin* 34(2), pp. 159-170. doi: 10.1177/0146167207310022.

Ford, T. et al. 2014. Not all groups are equal: Differential vulnerability of social groups to the prejudice-releasing effects of disparagement humor. *Group Processes & Intergroup Relations* 17(2), pp. 178-199. doi: 10.1177/1368430213502558.

Foucault, M. 1995. *Discipline and punish*. New York: Vintage Books.

Fraser, B. 1990. Perspectives on Politeness. *Journal of Pragmatics* 14(2), pp. 219-236. doi: 10.1016/0378-2166(90)90081-n.

Fredheim, R. et al. 2015. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2591299.

Fuchs, C. 2011. Web 2.0, Prosumption, and Surveillance. *Surveillance & Society* 8(3), pp. 288-309. doi: 10.24908/ss.v8i3.4165.

Gab Social. 2020. Available at: https://www.gab.com/ [Accessed: 8 June 2020].

Garcés-Conejos Blitvich, P. 2010. Introduction: The status-quo and quo vadis of impoliteness research. *Intercultural Pragmatics* 7(4)

Garcés-Conejos Blitvich, P. and Bou-Franch, P. 2019. Introduction to Analyzing Digital Discourses: New Insights and Future Directions. In: Bou-Franch, P. and Garcés-Conejos Blitvich, P. ed. *Analyzing Digital Discourse: New Insights and Future Directions*. 1st ed. Cham: Palgrave Macmillan

Garfinkle, H. 1967. *Studies in Ethnomethodology*. New Jersey: Prentice-Hall Inc.

Garland, D. 2000. The Culture of High Crime Societies. *British Journal of Criminology* 40(3), pp. 347-375. doi: 10.1093/bjc/40.3.347.

Garland, D. 2001. *The Culture of Control: Crime and Social Order in Contemporary Society*. 1st ed. Chicago: The University of Chicago Press.

Garland, J. and Chakraborti, N. 2012. Divided by a common concept? Assessing the implications of different conceptualizations of hate crime in the European Union. *European Journal of Criminology* 9(1), pp. 38-51.

Garland, J. and Hodkinson, P. 2014. 'F**king Freak! What the Hell Do You Think You Look Like?': Experiences of Victimization Among Goths and Developing Notions of Hate Crime. *British Journal of Criminology* 54(4), pp. 613-631. doi: 10.1093/bjc/azu018.

Glaser, B. and Strauss, A. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Routledge.

Goffman, E. 1959. *The Presentation of Self in Everyday Life*. New York: Doubleday.

Goffman, E. 1961a. *Asylums*. Middlesex: Anchor Books.

Goffman, E. 1961b. *Encounters: Two Studies in the Sociology of Interaction*. 1st ed. Indianapolis: Bobbs-Merrill.

Goffman, E. 1963. *Stigma*. 1st ed. New Jersey: Prentice-Hall Inc.

Goffman, E. 1967. *Interaction ritual: essays on face-to-face interaction*. 1st ed. New York: Routledge.

Goffman, E. 1971. *Relations in Public: Microstudies of the Public Order*. 2nd ed. New York: Basic Books, Inc.

Graham, S. and Hardaker, C. 2017. (Im)politeness in Digital Communication. In: Culpeper, J. et al. ed. *The Palgrave Handbook of Linguistic (Im)politeness*. 1st ed. London: Palgrave Macmillan

Greenwalt, K. 1989. *Speech Crime & the Uses of Language*. Oxford: Oxford University Press.

Grice, H. 1975. Logic and Conversation. *Speech Acts* , pp. 41-58. doi: 10.1163/9789004368811_003.

Hate crime | The Crown Prosecution Service. 2021. Available at: https://www.cps.gov.uk/crime-info/hate-crime [Accessed: 16 June 2021].

Hateful conduct policy. 2020. Available at: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [Accessed: 9 June 2020].

Hate speech policy - YouTube Help. 2020. Available at: https://support.google.com/youtube/answer/2801939?hl=en [Accessed: 8 June 2020].

Haugh, M. 2010. Intercultural (im)politeness and the micro-macro issue. In: Trosborg, A. ed. *Pragmatics across languages and cultures*. Berlin/New York: Walter de Gruyter

Haugh, M. 2014. Jocular Mockery as Interactional Practice in Everyday Anglo-Australian Conversation. *Australian Journal of Linguistics* 34(1), pp. 76-99. doi: 10.1080/07268602.2014.875456.

Haugh, M. 2016. "Just kidding": Teasing and claims to non-serious intent. *Journal of Pragmatics* 95, pp. 120-136. doi: 10.1016/j.pragma.2015.12.004.

Hernández, L. and Mendoza, F. 2002. Grounding, semantic motivation, and conceptual interaction in indirect directive speech acts. *Journal of Pragmatics* 34(3), pp. 259-284. doi: 10.1016/s0378-2166(02)80002-9.

Herring, S. 1996. *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. 1st ed. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Herring, S. 2004. Computer-mediated discourse analysis: An approach to researching online behavior. In: Barab, S. et al. ed. *Designing for virtual communities in the service of learning*. 1st ed. New York: Cambridge University Press, pp. 338-376.

Herring, S. and Androutsopoulos, J. 2015. Computer-Mediated Discourse 2.0. In: Tannen, D. et al. ed. *The Handbook of Discourse Analysis*. 2nd ed. Chichester: Wiley-Blackwell

Hinds, L. and Grabosky, P. 2010. Responsibilisation Revisited: From Concept to Attribution in Crime Control. *Security Journal* 23(2), pp. 95-113. doi: 10.1057/palgrave.sj.8350089.

Hodges, A. 2015. Intertextuality in Discourse. In: Tannen, D. et al. ed. *The Handbook of Discourse Analysis*. 2nd ed. Chichester: Wiley

Home Office 2012. *Hate crime action plan: Challenge it, Report it, Stop it*. Gov.UK.

Home Office 2015. *Hate Crime, England and Wales, 2014/15*. London: Home Office.

Hope, T. 2009. The illusion of control: A response to Professor Sherman. *Criminology and Criminal Justice* 9(2), pp. 125-134.

House, A. et al. 2011. Interpersonal trauma and discriminatory events as predictors of suicidal and nonsuicidal self-injury in gay, lesbian, bisexual, and transgender persons. *Traumatology* 17(2), pp. 75-85. doi: 10.1177/1534765610395621.

Hubbard, L. 2021. *Hate Crime Report 2021*. GALOP. Available at: https://galop.org.uk/wp-content/uploads/2021/06/Galop-Hate-Crime-Report-2021-1.pdf [Accessed: 21 July 2021].

Hymes, D. 1962. The Ethnography of Speaking. In: Gladwin, T. and Sturtevant, W. ed. *Anthropology and Human Behavior*. Washington DC: The Anthropological Society of Washington

Hymes, D. 1972. Models of the interaction of language and social life. In: Gumperz, J. and Hymes, D. ed. *Directions in Sociolinguistics*. New York: Holt, Rinehart and Winston

Hymes, D. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. Philadelphia: University of Pennsylvania Press.

Jaworski, A. and Coupland, N. 1999. Introduction: Perspectives on Discourse Analysis. In: Jaworski, A. and Coupland, N. ed. *The Discourse Reader*. 2nd ed. Oxon: Routledge

Jones, R. et al. 2015. Introduction: discourse analysis and digital practices. In: Jones, R. et al. ed. *Discourse and Digital Practices: Doing discourse analysis in the digital age*. 1st ed. Oxon: Routledge

Khan, M. 2017. Social media engagement: What motivates user participation and consumption on YouTube?. *Computers in Human Behavior* 66, pp. 236-247.

Kiesler, S. et al. 1984. Social psychological aspects of computer-mediated communication. *American Psychologist* 39(10), pp. 1123-1134. doi: 10.1037/0003-066x.39.10.1123.

Koh, H. 2003. On American Exceptionalism. *Stanford Law Review* 55(5), pp. 1479-1527.

Kohut, H. 1977. *The Restoration of the Self*. New York: International Universities Press.

Laaksonen, S. et al. 2020. The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data* 3

Levin, J. and McDevitt, J. 1993. *Hate crimes*. New York: Plenum Press.

Labov, W. 1997. Rules for Ritual Insults. In: Coupland, N. and Jaworski, A. ed. *Sociolinguistics: A Reader*. New York: Macmillan Education

Lakoff, G. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic* 2(4). doi: 10.1007/bf00262952.

Lakoff, G. 1987. *Women, fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

LeBlanc, T. 2010. Impolitness as a Model for Virtual Speech Community Building. In: Taiwo, R. ed. *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*. Hershey: Information Science Reference

Leech, G. 1983. *Principles of pragmatics*. London: Longman.

*Legal Aid, Sentencing and Punishment of Offenders Act 2012* London. c.10.

Levin, B. 2002. Cyberhate: A Legal and Historical Analysis of Extremists' Use of Computer Networks in America. *American Behavioral Scientist* 45(6), pp. 958-988. doi: 10.1177/0002764202045006004.

Lewis, H. 1971. *Shame and Guilt in Neurosis*. 1st ed. New York: International Universities Press.

Liu, R. 1986. *The politeness principle in A Dream of Red Mansions*. Unpublished M.Phil, Lancaster University.

Locher, M. and Watts, R. 2005. Politeness Theory and Relational Work. *Journal of Politeness Research. Language, Behaviour, Culture* 1(1). doi: 10.1515/jplr.2005.1.1.9.

Loftus, E. et al. 1987. Some facts about "weapon focus..." *Law and Human Behavior* 11(1), pp. 55-62. doi: 10.1007/bf01044839.

Lomash, E. et al. 2019. “A Whole Bunch of Love the Sinner Hate the Sin”: LGBTQ Microaggressions Experienced in Religious and Spiritual Context. *Journal of Homosexuality* 66(10), pp. 1495-1511. doi: 10.1080/00918369.2018.1542204.

López, I. 2014. *Dog Whistle Politics*. New York: Oxford University Press.

Mair, M. and Sharrock, W. 2021. Action, meaning and understanding: seeing sociologically with Harvey Sacks. In: Smith, R. et al. ed. *On Sacks: Methodology, Materials, and Inspirations*. Oxon: Routledge

*Malicious Communications Act 1988* London: The National Archives. c.27.

Marwick, A. and boyd, d. 2010. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1), pp. 114-133. doi: 10.1177/1461444810365313.

Mathiesen, T. 1997. The Viewer Society: Michel Foucault's `Panopticon' Revisited. *Theoretical Criminology*, 1(2), pp.215-234.

McDevitt, J. et al. 2002. Hate Crime Offenders: An Expanded Typology. *Journal of Social Issues* 58(2), pp. 303-317. doi: 10.1111/1540-4560.00262.

McKee, R. 2013. Ethical issues in using social media for health and health care research. *Health Policy* 110, pp. 298-301.

Mead, G. 1934. *Mind, Self, and Society: From the standpoint of a social behaviorist*. 1st ed. Chicago: The University of Chicago Press.

Mehlman, R. and Snyder, C. 1985. Excuse Theory: A Test of the Self-Protective Role of Attributions. *Journal of Personality and Social Psychology* 49(4), pp. 994-1001. doi: 10.1037/0022-3514.49.4.994.

Miller, C. 1984. Genre as social action. *Quarterly Journal of Speech* 70(2), pp. 151-167. doi: 10.1080/00335638409383686.

Mooney, J. 2000. *Gender, violence and the social order*. London: Macmillan Press.

Murphy, S. 2017. Humor Orgies as Ritual Insult: Putdowns and Solidarity Maintenance in a Corner Donut Shop. *Journal of Contemporary Ethnography* 46(1), pp. 108-132. doi: 10.1177/0891241615605218.

Nissenbaum, H. 2010. *Privacy in Context: Technology, Policy and the Integrity of Social Life*. Stanford: Stanford University Press.

Novaes, C. 2016. Reductio ad absurdum from a dialogical perspective. *Philosophical Studies* 173(10), pp. 2605-2628. doi: 10.1007/s11098-016-0667-6.

Nowicka, M. 2018. “I don’t mean to sound racist but … ” Transforming racism in transnational Europe. *Ethnic and Racial Studies* 41(5), pp. 824-841. doi: 10.1080/01419870.2017.1302093.

Office for National Statistics 2012. *Crime Survey For England And Wales (CSEW)*. Office for National Statistics.

Ogiermann, E. 2009. *On apologising in negative and positive politeness cultures*. Amsterdam: John Benjamins.

O'Reilly, T. 2005. Web 2.0: Compact Definition?. *Radar: Insight, Analysis, and Research about Emerging Technologies* . Available at: http://radar.oreilly.com/2005/10/web-20-compact-definition.html [Accessed: 16 June 2021].

Paccagnella, L. 1997. Getting the Seats of Your Pants Dirty: Strategies for Ethnographic Research on Virtual Communities. *Journal of Computer-Mediated Communication* 3(1). doi: 10.1111/j.1083-6101.1997.tb00065.x.

Panther, K. and Thornburg, L. 1998. A cognitive approach to inferencing in conversation. *Journal of Pragmatics* 30(6), pp. 755-769. doi: 10.1016/s0378-2166(98)00028-9.

Papacharissi, Z. 2004. Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society* 6(2), pp. 259-283. doi: 10.1177/1461444804041444.

Peeters, R. 2013. Responsibilisation on Government's Terms: New Welfare and the Governance of Responsibility and Solidarity. *Social Policy and Society* 12(4), pp. 583-595.

Perry, B. and Olsson, P. 2009. Cyberhate: The Globalization of Hate. *Information & Communications Technology Law* 18(2), pp. 185-199. doi: 10.1080/13600830902814984.

Phillips, J. and Bartlett, J., 2018. *Should anonymous social media accounts be banned?*. [online] The Guardian. Available at:

https://www.theguardian.com/media/2018/sep/30/social-media-anonymity-ban-debate-trolls-abuse--jess-phillips-jamie-bartlett [Accessed 2 June 2021].

Phillips, C. and Bowling, B. 2012. Ethnicities, racism, crime and criminal justice. In: Maguire, M. et al. ed. *The Oxford Handbook of Criminology*. Oxford: Oxford University Press

Planchenault, G. 2010. Virtual community and politeness: The use of female markers of identity and solidarity in a transvestites' website. *Journal of Politeness Research. Language, Behaviour, Culture* 6(1). doi: 10.1515/jplr.2010.005.

Pomerantz, A. 1984. Agreeing and disagreeing with assessment: Some features of preferred/dispreferred turn shapes. In: Atkinson, J. and Heritage, J. ed. *Structure of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press

Potter, J. 1996. *Representing Reality*. London: Sage.

*Protection from Harassment Act 1997* London: The National Archives. c.40.

*Public Order Act 1986* London: UK Government. c.64.

Rainie, L., Anderson, J. and Albright, J., 2017. *The Future of Free Speech, Trolls, Anonymity and Fake News Online*. [online] Pew Research Center: Internet, Science & Tech. Available at: https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/ [Accessed 2 June 2021].

*Recommendation No. R (97) 20* Council of Europe.

Reisigl, M. and Wodak, R. 2001. *The semiotics of racism*. Wien: Passagen.

Reisigl, M. and Wodak, R. 2016. The discourse-historical approach. In: Wodak, R. and Meyer, M. ed. *Methods of critical discourse analysis*. 3rd ed. London: Sage

Richards, R. and Calvert, C. 2000. Counterspeech 2000: A New Look at the Old Remedy for "Bad" Speech. *BYU Law Review* 2000(2), pp. 553-586.

Richardson, B. et al. 2014. Language style matching and police interrogation outcomes. *Law and Human Behavior* 38(4), pp. 357-366. doi: 10.1037/lhb0000077.

Ritchie, K. 2017. Social Identity, Indexicality, and the Appropriation of Slurs. *Croatian Journal of Philosophy* 17, pp. 155-180.

Roach, L. 2015. *A comparative analysis of Online and Print Media during the Charlie Hebdo attack.* MSc, Cardiff University.

Roach, L. 2016. *Action and Reaction: An Analysis of Hate Speech and Counter Speech on Social Media*. MSc, Cardiff University.

Rousseau, J. 1994. *Discourse on Political Economy and The Social Contract*. New York: Oxford University Press.

Sacks, H. 1972. An initial investigation of the usability of conversational data for doing sociology. In: Sudnow, D. ed. *Studies in Language and Social Interaction*. New York: Free Press, pp. 31-74.

Sacks, H. and Jefferson, G. 1995. *Lectures on conversation*. Oxford: Blackwell.

Scheff, T. 1987. The Shame-Rage Spiral: A Case Study of an Interminable Quarrel. In: Lewis, H. ed. *The Role of Shame in Symptom Formation*. 1st ed. Hillsdale, N.J.: L. Erlbaum Associates., pp. 109-149.

Scheff, T. and Retzinger, S. 2001. *Emotions and violence: shame and rage in destructive conflicts*. Lincoln: Backinprint.com.

Schegloff, E. 2006. *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.

Schieb, C. and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook. In: *66th ICA Annual Conference*. Fukuoka: International Communication Association

Shearing, C. and Wood, J. 2003. Nodal Governance, Democracy, and the New 'Denizens'. *Journal of Law and Society* 30(3), pp. 400-419. doi: 10.1111/1467-6478.00263.

Sherman, L. 2009. Evidence and Liberty: The Promise of Experimental Criminology. *Criminology & Criminal Justice* 9(1), pp. 5-28. doi: 10.1177/1748895808099178.

Sherzer, J. 1983. *Kuna ways of speaking*. Austin: University of Texas Press.

Sifianou, M. and Bella, S. 2019. Twitter, Politeness, Self-Presentation. In: Bou-Franch, P. and Blitvitch, P. ed. *Analyzing Digital Discourse: New Insights and Future Directions*. 1st ed. Cham: Palgrave Macmillan

Silverman, D. 1998. *Harvey Sacks: Social Science and Conversation Analysis*. New York: Oxford University Press.

Slayden, D. and Whillock, R. 1995. *Hate speech*. London: Sage.

Smith, R. 2017. Membership categorisation, category-relevant spaces, and perception-in-action: The case of disputes between cyclists and drivers. *Journal of Pragmatics* 118, pp. 120-133. doi: 10.1016/j.pragma.2017.05.007.

Stenovec, T. 2017. Those dots you see in iMessage are more complicated than you think [Online]. Available at: http://uk.businessinsider.com/the-imessage-dots-explained-2016-1?r=US&IR=T [Accessed: 7 December 2017].

Stikkers, K. 2014. " . . . But I'm Not Racist": Toward a Pragmatic Conception of "Racism". *The Pluralist* 9(3), pp. 1-17. doi: 10.5406/pluralist.9.3.0001.

Stop Hate UK 2020. *Stop Hate UK Annual Report 2019-20*. Stop Hate UK. Available at: https://www.stophateuk.org/wp-content/uploads/2021/01/Annual-report-19_20-1.pdf [Accessed: 21 July 2021].

Stubbs, M. 2015. Computer-Assisted Methods of Analyzing Textual and Intertextual Competence. In: Tannen, D. et al. ed. *The Handbook of Discourse Analysis*. 2nd ed. Chichester: Wiley

Swales, J. 1990. *Genre analysis: English in academic and research settings*. 1st ed. Cambridge: Cambridge University Press.

Swales, J. 2004. *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Sykes, G. and Matza, D. 1957. Techniques of Neutralization: A Theory of Delinquency. *American Sociological Review* 22(6), pp. 664-670. doi: 10.2307/2089195.

Tedeschi, J. and Riess, M. 1981. Verbal strategies in impression management. In: Antaki, C. ed. *The psychology of ordinary explantions of social behavior*. London: Academic Press

TellMAMA 2019. *Tell MAMA Annual Report 2018: Normalising Hatred*. TellMAMA. Available at: https://tellmamauk.org/tell-mama-annual-report-2018-_-normalising-hate/ [Accessed: 21 July 2021].

Terrion, J. and Ashforth, B. 2002. From 'I' to 'we': The role of putdown humour and identity in the development of a temporary group. *Human Relations* 55(1), pp. 55-88.

Thurlow, C. 2017. Digital discourse: Locating language in new/social media. In: Burgess, J. et al. ed. *Handbook of Social Media*. New York: Sage

Tilley, N. 2009. Sherman vs Sherman: Realism vs rhetoric. *Criminology & Criminal Justice* 9(2), pp. 135-144. doi: 10.1177/1748895809102549.

Townsend, L. and Wallace, C. 2017. The Ethics of Using Social Media Data in Research: A New Framework. In: Woodfield, K. ed. *The Ethics of Online Research: Volume 2*. Bingley: Emerald Publishing Limited

Twitter 2017. About verified accounts [Online]. Available at: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts [Accessed: 7 December 2017].

Vásquez, C. 2015. Intertextuality and interdiscursivity in online consumer reviews. In: Jones, R. et al. ed. *Discourse and Digital Practices: Doing discourse analysis in the digital age*. 1st ed. New York: Routledge

van Leeuwen, T. 1993. Genre and Field in Critical Discourse Analysis. *Discourse & Society* 4(2), pp. 193-225. doi: 10.1177/0957926593004002004.

van Leeuwen, T. 2007. Legitimation in discourse and communication. *Discourse & Communication* 1(1), pp. 91-112. doi: 10.1177/1750481307071986.

Velody, I. and Williams, R. 1998. Introduction. In: Velody, I. and Williams, R. ed. *The Politics of Constructionism*. London: Sage

von Wallpach, S. et al. 2017. Performing identities: Processes of brand and stakeholder identity co-construction. *Journal of Business Research* 70, pp. 443-452. doi: 10.1016/j.jbusres.2016.06.021.

Walters, M. et al. 2017. *Hate Crime and the Legal Process: Options for Law Reform*. Sussex: University of Sussex.

Wang, Y. et al. 2014. Linguistic Adaptation in Conversation Threads: Analyzing Alignment in Online Health Communities. In: *ACL Workshop on Cognitive Modeling and Computational Linguistics*. Baltimore: Association for Computational Linguistics, pp. 55-62. Available at: http://cani.ist.psu.edu/publication/LinguisticAdaptationCMCL2014.pdf [Accessed: 22 June 2021].

Wartenberg, T. 1990. *The Forms of Power: From Domination to Transformation*. Philadelphia: Temple University Press.

Wasserman, H. 2003. Symbolic Counter-Speech. *Williams and Mary Bill of Rights Journal* 12(2)

Webb, H. et al. 2016. Digital Wildfires: Propagation, Verification, Regulation, and Responsible Innovation. *ACM Transactions on Information Systems* 34(3), pp. 1-23. Available at: https://dl.acm.org/doi/10.1145/2893478 [Accessed: 16 June 2021].

Welsh Government 2014. *Tackling Hate Crimes and Incidents - Framework for Action*.

West, L. and Trester, A. 2013. Facework on Facebook. In: Tannen, D. and Trester, A. ed. *Discourse 2.0: Language and New Media*. Washington: Georgetown University Press

Wilkinson, D. 2001. Violent events and social identity: Specifying the relationship between respect and masculinity in inner-city youth violence. In: Kinney, D. ed. *Sociological Studies of Children and Youth Vol. 8*. Bingley: Emerald Group Publishing Limited

Williams, M. et al. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology* 51(6), pp. 1149-1168.

Williams, M. et al. 2017b. Users' View of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm. In: Woodfield, K. ed. *The Ethics of Online Research: Volume 2*. Bingley: Emerald Publishing Limited

Williams, M. 2019. *Hatred behind the screens: A report on the rise of online hate speech.*. Mishcon de Reya. Available at: https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf [Accessed: 15 June 2021].

Williams, M. 2021. *The Science of Hate: How prejudice becomes hate and what we can do to stop it*. 1st ed. London: Faber & Faber Ltd.

Wilson, J. and Kelling, G. 1982. Broken Windows: The police and neighbourhood safety. *The Atlantic* . Available at: https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/ [Accessed: 16 June 2021].

Wooffitt, R. 2005. *Conversation Analysis and Discourse Analysis: A Comparative and Critical Introduction*. London: Sage.

Wright, L. et al. 2017. Vectors for Counterspeech on Twitter. In: *Workshop on Abusive Language Online*. Vancouver: Association for Computational Linguistics, pp. 57-62.

Zidjaly, N. 2010. Intertextuality and Constructing Islamic Identities Online. In: Taiwo, R. ed. *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*. 1st ed. Hershy: Information Science Reference