1

Building and Analysing an Online Hate Speech Corpus: The NETLANG Experience and Beyond

Isabel Ermida

1 Introduction

Every edited volume is a joint enterprise, resulting from the concerted efforts of a set of people with common research interests who happen to have crossed paths at a certain, fruitful, point in time. This book is no exception, but its group dynamics go deeper than the occasional CFP. Half the following chapters are by researchers who interacted closely for four years, the distance and the pandemic notwithstanding, as members of an

I. Ermida (✉)
University of Minho, Braga, Portugal
e-mail: iermida@elach.uminho.pt

I. Ermida (ed.), *Hate Speech in Social Media*,
https://doi.org/10.1007/978-3-031-38248-2_1

international project, "NETLANG".[1] The other half is by scholars who flew many miles to meet the team in Portugal, during a heat wave that did not manage to prevent the project's final conference from being a prosperous forum of discussion. The topic, granted, is dark—but the dialogue around it was bright and inspiring, so much so that the following pages are a mosaic of different shafts of light, different views, and different perspectives, each unveiling a special facet of hate speech.

The "darkness" of hate speech does not confine it to the dark web (Gehl, 2016). Instead, hate speech promenades along mainstream social media in broad daylight. "Social media", by the way, is here understood as "Internet-based channels of masspersonal communication" which "derive value primarily from user-generated content" and allow users to "opportunistically interact and selectively self-present" (Carr & Hayes, 2015: 8). No special web browsers or tailored routers are needed to access hate speech: a connection to the Internet is all it takes. That is why surfing the web, handy and affordable as it has become, may nowadays be a perilous pastime. Online platforms admittedly keep the world open to millions, and allow communicators to creatively interact on a global scale, but they also set traps to the unwary user. Chauvinism, discrimination, oppression—all have found their way into online forums, collecting victims at the same rate as followers, and feeding ever greater polarisation, inequality, and radicalisation in society.

Whenever prejudiced content is voiced online, an indeterminate, but potentially vast, range of people are subject to its pernicious impact. First and foremost, the targets, who directly bear the full force of the attack, both individually and collectively, are thence humiliated, offended, dehumanised, "othered", and, by means of fear, silenced (e.g. Benesch, 2014). Gradually and surreptitiously, hate speech contributes to isolate,

[1] Funded by the Portuguese Foundation for Science and Technology, the NETLANG project integrated linguists from five different European countries: besides Portugal, Czech Republic, Estonia, Finland, and Poland. It also integrated researchers from other areas besides linguistics: computer scientists, psychologists, law and education scholars. The project's full title, which bears the initial adoption of a term ("cyberbullying") which later came to be overshadowed and definitely replaced by "hate speech", is "The Language of Cyberbullying: Forms and Mechanisms of Online Prejudice and Discrimination in Annotated Comparable Corpora of Portuguese and English" (ref. PTDC/LLT-LIN/29304/2017).

See https://sites.google.com/site/projectnetlang/team

marginalise, disparage, and demonise vulnerable individuals and the communities they represent, causing them to see their rights jeopardised and their public image soiled. But the bystanders, or interlocutors, are also (indirectly) affected, and play a central, perlocutionary role in the hate speech act: they absorb, more or less distractedly, the biased social meanings fed to them, and may be persuaded to replicate them. In other words, they are the targets of incitement, and simultaneously the recipients, or decoders, of hatred aimed at third parties (Assimakopoulos, 2020; O'Driscoll, 2020). If they yield to such a malign influence, they may end up reproducing it and becoming hate emissaries themselves, or "hate recruits", as Langton (2018) puts it, possibly engaging in active discrimination and even actual, physical violence. Here lies a major danger of online hate speech: the contamination effect, due to its "toxicity" (see Konikoff, 2021) and potential for "pollution" (Nagle, 2009). Haters are, actually, agents of ideological contagion at a global scale, lurking, with perceived impunity, behind the anonymous or pseudonymous hideout that a mere login provides (Woods & Ruscher, 2021), while the online permanence of their written commentary allows it to re-emerge and recapture support over time. The disease they spread is prejudice; the side-effects are heightened bigotry and intolerance.

From a control and regulation perspective, the focus lies in employing ever more intelligent hate detection algorithms and establishing anti-discrimination policies, which necessarily have different legal expressions across countries (Banks, 2010). Public entities, from governments and political organisations to the digital companies responsible for supplying the social media services, all have struggled to keep hate speech under control, all the while respecting the overarching right to freedom of expression. This is a very challenging compromise, of which haters are well aware—and due to which they skilfully dodge the moderation obstacles by re-inventing themselves, constantly finding renewed, imaginative ways to convey their hateful contents. The centrality of language to understanding and monitoring hate speech is therefore undeniable: it is by manipulating it that haters accomplish their agendas, and it is by scrutinising it that moderators can spot and counter such agendas.

Academics and non-academics alike have thus concentrated on capturing the ever evasive language used in hate speech. The NETLANG

project team, just like the other researchers that joined them at the final conference, tried to tackle this very challenge. All through the duration of the project, as well as through the three days of its epilogue, hate-speech language was laid on a stretcher, stripped of its many outfits and disguises, and dissected to the bone. Many questions came out unanswered, alas, but many others were raised. Before turning to the analytical outcomes of the NETLANG project, together with the analyses of hate speech provided by the external contributors to this collection, a synopsis of existing research into the language of hate is in order.

2 Linguistic Approaches to Hate Speech

Linguistic research into hate speech has grown steadily over the past few years, covering different levels of linguistic analysis and springing from various theoretical frameworks. This section offers an overview of such scholarship, which encompasses an array of phenomena, ranging from overt, explicit forms of hatred to covert, implicit strategies to convey prejudice and discrimination, most of which surface, at one time or another, throughout the present book.

The lexical features of hateful language, prominent and manifest as they are, have naturally attracted much academic attention. Slurs and taboo words used in prejudiced discourse have been extensively analysed by such authors as Stokoe and Edwards (2007), who examine racial insults in police interrogations; Williamson (2009), who discusses the semantics of pejoratives and the merits of inferentialism, a non-truth-conditional approach to slurs; and Hedger (2013), who looks at the semantic grounding of derogatory racial epithets. Similarly, Anderson and Lepore (2013) investigate the fluctuations in the offensive potential of insults, whereas Nunberg (2018) reflects upon the dual—both descriptive and evaluative—nature of slurs.

The lexis of hate has also been explored in terms of keyness, frequency, and collocation, with the help of computational tools (e.g. Waseem & Hovy, 2016; Fersini et al., 2018; Zampieri et al., 2019). Automatic detection models based on lexical elements depart from the assumption that the presence of certain negative words (such as invectives and disparaging

nouns and adjectives) can be used as a feature for classifying a text as hateful (Schmidt & Wiegand, 2017). This task requires lexical resources that contain such predictive elements, like ontologies and dictionaries. Some of these dictionaries focus on content words, such as insults and swearwords (see "Noswearing", by Liu & Forss, 2015), profane words, including acronyms and abbreviations (Dadvar et al., 2012), and "label specific features", such as forms of verbal abuse and widespread stereotypes (see the Ortony Lexicon, by Dinakar et al., 2011), all of which can be searched as potentially hateful keywords. Despite the popularity of keyness approaches (the NETLANG project also resorts to them—see below), they have been revealed to be problematic (e.g. Scott, 2010). As MacAvaney et al. (2019) rightly remark, keyword-based approaches may have high precision (i.e. high percentage of relevant hits from the set rated as hate speech) but low recall (i.e. low percentage of relevant hits from the global dataset). This implies that a system relying only on keywords would not identify hateful content that does not use them (false negatives). Conversely, it would also create false positives, since certain potentially pejorative keywords—like "trash" or "swine"—may be used in neutral, "clean" passages.

Crucially, keyword-based approaches cannot handle figurative, implicit or nuanced language, which skilfully phrases hate in disguised ways. For instance, comments such as "Hold on, where did I put my harpoon?", or onomatopoeias like "Oink!", produced in response to a text about an overweight model, sport no hateful keyword whatsoever—and yet express body shaming as regards obese people, a type of discrimination against a vulnerable group (Ermida, 2014). One last hitch overshadowing keyword lexical approaches is the deliberate obfuscation of words and phrases to evade automatic detection, as is the case of racist misspellings like "ni9 9er", "whoopiuglyniggerratgoldberg" and "JOOZ" (Nobata et al., 2016).

Many other linguistic approaches to hate speech stem from the realm of pragmatics, assuming an integrative view of hate speech in its social and interactional context. Critical Discourse Analysis (CDA) informs most scholarship in this regard (see Chovanec, and Ruzaitė, this volume). After all, as Assimakopoulos et al. (2020) point out, the CDA analytical paradigm typically looks at the linguistic expression of ideologically charged attitudes, especially as regards discrimination. Brindle (2016),

for instance, has combined CDA with corpus linguistics to study frequency, collocations, and concordances expressing gender prejudice in a white supremacist online forum. Sharifi et al. (2017), likewise, use a critical discourse analytic approach to tackle the prejudiced construction of Islam in Western talk shows. Along similar lines, Esposito and Zollo (2021) apply a so-called Social Media Critical Discourse Studies perspective to the analysis of online misogyny, covering such textual phenomena as animal metaphors, body shaming insults involving synaesthesia, and slurs concerning purported gender identity, while Raffone (2022) combines CDA with a social-semiotic framework to analyse disability hate speech on TikTok. In particular, the sociocognitive paradigm within Critical Discourse Analysis has been very productive: see, for instance, Đorđević's (2019) sociocognitive approach to readers' comments on Serbian news websites, and Sirulhaq et al.'s (2023) comprehensive survey of a sociocognitive-CDA methodology in hate speech studies. Rather tellingly, Van Dijk (2005), an exponent of sociocognitive CDA, has dedicated his life's research to studying prejudice, especially racist prejudice, in public discourse, and has recently inflected his attention to anti-racist discourse, one form of counterspeech (cf. Van Dijk, 2021).

Cognitive Linguistics is another theoretical pool with pragmatic import for hate speech inquiry, especially as applied to conceptual metaphor (see Laineste & Chlopicki, this volume). Two examples are Musolff (2017) and Pražmo (2020), who look into the dehumanising metaphors that are used, respectively, in UK anti-immigrant debates on social media, and misogynistic prejudice in the InCel online community. Cognitive Linguistics has also been productive in exploring the concept of mental spaces in hate speech. Lewis (2012) elaborates on how mental spaces, blending, and related cognitive domains underlie the emotional construction of offense in hate speech, while Raj and Usman (2021) also resort to the notion of mental spaces in conceptualising hate speech on Facebook, in terms of base spaces and space builders that affect the perception or interpretation of hateful language.

Speech Act Theory is a very fruitful framework for the analysis of hate speech. Austin and Searle's account of intentionality, in particular, has proved important for feminist and anti-racist views of language, which have asked for regulatory intervention against hate speech. In the

1990s, a raging debate involved such scholars as Mackinnon (1993), Langton (1993), Matsuda et al. (1993), Hornsby (1995), and Lederer and Delgado (1995), all of whom regard hate speech as conveying a linguistically identifiable intent (illocutionary force) to cause harm, namely to subordinate and marginalise members of an oppressed group, and hence as a practice that should be subject to restriction. In particular, Langton's (1993) "Speech Acts and Unspeakable Acts" applies speech act insight into the analysis of women's oppression through pornography. A highly polemical response was Butler's (1997) *Excitable Speech: A Politics of the Performative*, which stresses the "open-ended" nature of speech acts and the difficulty in attributing intentional blame, and advises against hate speech censorship on the grounds that it could paradoxically "silence" the victims, who otherwise could be roused to defy hate speech by "resignifying" and "restaging" it (on a counter-response to Butler, see Schwartzman, 2002).

The dispute around the performativity of hate speech and the accountability of the illocutionary expression of discrimination and prejudice has regularly re-emerged. Tsesis (2009: 518), for instance, speaking from a legal standpoint, points out that "speech acts that rely on culturally recognized images of subordination" are not simply "the sentiments of a single person"; instead, they rely on "the symbolic efficacy of group slogans to express acceptable conduct toward a named class of individuals" (see also Gelber, 2017; and Weston, 2022, on whether or not speech can "perform" regulable action). Along similar lines, Carney (2014), adopts a Searlean framework to defend that examining the speech acts of a verbal exchange does make it possible to assess whether the speaker's words are either hurtful or harmful (see also Özarslan, 2014, on applying speech act theory to hate speech studies online, in the era of Web 2.0). More recently, O'Driscoll (2020) has put forth interesting Searlean considerations, trying to theorise vaguer, indeterminate forms of intentionality and incitement, whereas MacDonald and Lorenzo-Dus (2020) have offered a Speech Act Theory discussion of persuasion with regard to terrorism incitement. Significantly, Assimakopoulos (2020) has examined how incitement to discriminatory hatred relates to illocution and perlocution, and he proposes a reworked Searlean notion of felicity conditions.

Impoliteness Studies also rank amongst the most prevalent approaches to the language of hate (see Faria, and Ruzaitė, this volume). Actually, a good few years before the term "hate speech" became popular in linguistic research, a number of forerunning studies were already approaching hateful language in computer-mediated communication (CMC) from the perspective of impolite phenomena. In 2011, for instance, Lorenzo-Dus et al. examined the impoliteness of prejudiced online polylogues, and in 2014 Lange discussed perceptions of impoliteness and inappropriateness in face of speech by haters, as opposed to "ranters", on YouTube, while Carney (2014) also applied impoliteness concepts to tell hurtful from harmful content in public media. In the same year, Ermida's (2014) analysis of sexism and body shaming in online newspaper comment boards also resorted to the impoliteness framework, as did her study of classist prejudice against the poor and unemployed in the *Daily Mail* message forums four years later (Ermida, 2018). The application of impoliteness studies to online hate speech became gradually established in the second part of the decade. Hardaker and McGlashan's (2016) influential article on hate against women on Twitter is a central example of this trend, which Culpeper et al. (2017) crucially epitomise, by focusing on extreme religious hate speech assuming criminal contours (e.g. threat). Two other cases deserving mention are Kienpointer's (2018) analysis of what he labels "destructively impolite utterances" in hate speech across a variety of online genres (discussion forums, blogs, social media, tweets, homepages), and Carr et al.'s (2020) proposal of an impoliteness annotation scheme to improve the precision of hate speech detection. In his recent programmatic article, Culpeper (2021) compares and contrasts the two phenomena, impoliteness and hate speech, in a metapragmatic, first-order approach, and concludes that users employ the qualifier "hateful", rather than "impolite", to characterise more extreme behaviours with associations of prejudice.

Argumentation Studies—which historically stem from the combined fields of logic, dialectic, and rhetoric—are yet another relevant theoretical source for the linguistic analysis of hate speech, given its applications to civil debate, conversation, and persuasion (see Faria, and Ruzaitė, this volume). Examples are Burke et al. (2020), who examine argument and reasoning in Islamophobic and anti-Semitic discussions on Facebook,

Domínguez-Armas et al. (2023), who explore the argumentative functions, and the provocative effect, of mentioning ethnicity in prejudiced headlines, and Pettersson and Sakki (2023), who look into argumentation and polarisation in online debates around gender and radical-right populism.

Humour Studies are a characteristically interdisciplinary field of research, but the linguistic embedding of humour, together with its contextual dependency, make it a privileged object of, respectively, semantic and pragmatic attention. Linguistic approaches to humour in hate speech research are various and enlightening (see Laineste & Chlopicki, as well as Bick, and Ruzaitė, this volume), springing from the enlarged discussion of implicit forms of hate speech, and especially of the ways to downplay the assaulting force of the utterance and to avoid accountability. Vasilaki (2014), for instance, looks at name-calling in hate speech being mitigated through humour, while Godioli et al. (2022) also examine how humour can help defendants accused of hate speech in courts of law. Other authors investigate racist humour and discuss the limits of freedom of speech: Leskova (2016), for example, uses the pun "Black humour" to study hate speech against racial minorities in Europe; Trindade (2020) concentrates on "disparagement humour" being used to convey gendered racism on social media in Brazil; and Menon (2022: 5–6), focusing on white supremacist forums in America, scrutinises how humour "slips under the radar as a tool for spreading incendiary ideas" and how "the infrastructures of hate masquerade as harmless harbingers of humour". Further linguistic studies of humour in hate speech assume a computational stance and resort to Sentiment Analysis and Emotion Detection methods: Badlani et al. (2019) explore ways to disambiguate sentiment expressed in humour as used in hate speech, while Kazienko et al. (2023) also try to process subjective content in speech that is simultaneously humorous and hateful.

Although pragmatics has gained the upper hand in the critical landscape of hate language, grammatical studies have also been occasionally offered. As early as the 1980s, Van Dijk (1987) was discussing the use of nominalisation to replace SVO syntactic structures in racist discourse, where subject and verb would otherwise give away the agent of the problematic action: for instance, substituting the full clause "The police raid

killed Mrs Jarrett" (a Black woman who died from a heart attack during a police raid on her house), for the nominalised phrase "Mrs Jarrett's raid death" manages to not identify the police as the agents responsible for the raid, thus concealing their active role and accountability (Van Dijk, 1987: 78). Passivation is another case in point: Van Dijk (1987) also treats it as a strategy to divert the focus of an action from the actor to the recipient of the action. Similarly, Ehrlich (2001) studies the language used by men in rape trials, and she finds passive voice to be a strategy they employ to mitigate their responsibility (see also Ehrlich, 2014).

A very recent book, edited by Knoblock (2022), *The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanising Discourse*, supplies an exciting view into the role that grammatical elements play in conveying hate (and aggression). Some studies in the collection focus on morphological elements, such as English suffixes: Mattiello (2022) for instance, looks at the dysphemic use of the slang suffix –o (as in *lesbo*, *weirdo*, *thicko*), which adds a negative connotation to the stem, while Tarasova and Fajardo (2022) analyse the pejorative potential of the diminutive suffixes –ie/y (as in *brownie* and *blackie*). The exploitation of articles and pronouns for vilification purposes is the focus of other chapters: Lind and Nübling (2022) discuss how using German forms usually reserved for inanimate objects in reference to humans, namely women, is gravely disparaging, whereas Ohlson (2022) examines how switching from "he/she" to "it" is a powerful dehumanising strategy. Word formation processes inform yet other contributions to the book: Beliaeva (2022) investigates how lexical blends can be used both as humorous and as derogatory terms, whereas Korecky-Kröll and Dressler (2022) study expressive compound patterns including taboo nouns and deprecating adjectives. Finally, hateful discourse can also find an outlet in certain grammatical constructions, as is the case of those using imperative verbs (Bianchi, 2022) and syntactic patterns, such as the infamous "I am no racist but…" (Geyer et al., 2022).

Grammatical features have also informed automated models of hate speech detection, which try to consider the words in their syntactic environment. Gitari et al. (2015), for instance, have refined lexicon-based searches with co-occurring words in the sentence sequence, by resorting to part-of-speech (POS) tagging: if a noun like "Jews" appears as an

object of the verb "kill", the probability of hate speech to occur is higher and the intensity greater. In the same vein, Warner and Hirschberg (2012) have used three-element structures, or POS trigrams, to tackle similar issues, as in "DT Jewish NN" (e.g. "that Jewish b*stard"). Silva et al. (2016) have also applied syntactic structures to detect word relevance, as in "I<intensity><user intent><hate target>", e.g. "I f*cking hate Asian people".

Along similar lines, regular expressions, combining lexical elements with syntactic order, have been found to recur in the NETLANG corpus, and are good candidates to detect hate in other corpora as well. As Dias and Pereira show (this volume), some syntactic sequences produce good hate speech identification hits: this is the case of "(a | you) bunch of ((ADJ) + NN)", "if you think", "(x) people like you", "((if | whether) you) like it or not". Similarly, Ermida et al. (2023) have scraped the NETLANG corpus for the presence of adversative constructions signalling conflictual opinion exchanges, namely "I <to be/to have> <not/no> <NP/Adj> <but>", and the syntactic pattern does seem to be productive in detecting hate speech content. In the present collection, morphological and syntactic insight, for instance into the role of demonstratives (Aguiar & Barbosa), first-person verb forms (Biri et al.), and compounds (Bick), also provides a glimpse at the productivity and plasticity of specific structural elements in the hate speech sequence.

Crucially, and not surprisingly, computational linguistics has taken centre stage in research into the language of hate speech, given the automatic detection needs that digital companies experience when moderating user-generated content on their platforms. Corpus linguistics tools provide an important avenue for quantitative and automatic (often preliminary) treatments of large corpora such as the ones at hand. Many of the following chapters strive to combine quantitative methods with a qualitative examination of specific, manageable subsets, constructed by scraping massive databases. Only in this way can the contextual information surrounding the hateful comment, or embedded in its very cotextual structure, be reliably deciphered. Existing automatic detection tools—used, for instance, within Sentiment Analysis (or Opinion Mining) and Emotion Detection frameworks—manage to identify the negative polarity of a text and the probability that it expresses hateful

emotion (see Dias & Pereira, this volume). SentiWordNet (Baccianella et al., 2010), VADER (Valence Aware Dictionary and sEntiment Reasoner, see Hutto & Gilbert 2014), and JAMMIN3 (Argueta et al., 2016) are three such cases. Other popular corpus linguistics tools include Sketch Engine, which allows for a valuable range of exploratory inquiries and quantitative outputs, such as KWIC mappings and n-grams. In particular, Word Sketch analyses collocations, also in terms of grammatical relations, and produces statistical, rank-ordering results through logDice.

Most of the chapters in this collection exemplify the use of one or another of these tools. Importantly, the chapters that spring from the NETLANG project, just like those that result from similar projects (e.g. FRENK, XPEROPHS, Decoding Antisemistim), sport a computational linguistics design and assume, to a greater or lesser extent, a corpus-based methodology. So as to contextualise the present book and lay out the process that led to its creation, I will next introduce the construction of the NETLANG corpus and the methodological decisions that were made along the way.

3 Genesis of the Book: Construction and Preparation of an Online Hate Speech Corpus

The NETLANG project departed from the construction of a bilingual (English and Portuguese) corpus of potential hate speech, with a view to providing a large dataset basis for analysis, as well as material for future research. By the time the project reached its conclusion, a corpus of 50.5 million words had been built from scratch and made freely available online. This vast endeavour was motivated by the absence of a commonly accepted benchmark corpus for hate speech analysis, which has driven many other researchers to collect and label their own data from a variety of social media. Examples of such hate speech corpora are GermanTwitter (Ross et al., 2016), Hatebase (Davidson et al., 2017), Stormfront (Gilbert et al., 2018), XPEROHS (Baumgarten et al., 2019, see Chap. 6), and the FRENK corpus (Ljubešić et al., 2019, see Chap. 13). It should be noted

that, according to MacAvaney et al. (2019), existing hate speech corpora vary greatly in terms of size, scope, relevance, annotated data features, and annotated hate speech features. In the case of NETLANG, this time-consuming, complex data compilation task lasted, as expected, throughout the four years of the project. A brief summary of the team's work will frame this book, and shed light on the steps that led to the construction of the corpus, to its preparation in terms of data pre-processing, and to the various linguistic analyses for which it paved the way.

The bilingual nature of the corpus was initially intended to provide a balanced output of online verbal productions in the two languages (Portuguese, after all, is the eighth most spoken language worldwide, with c. 230 million native speakers, official language status in seven countries around four continents, and co-official status in another three). So, we tried to produce a similar number of "prompts" per language, i.e. of online posts, both verbal (texts) and audiovisual (videos), that might trigger the user-generated content—the comments—to analyse. But we soon realised, as the extraction process evolved, that the English subcorpus proved to be much more productive, both in terms of number of comments per post, and of number of words per comment. From the figure above, 43 million words overwhelmingly concern the English subcorpus alone. These were mainly extracted from the comment boards of YouTube, but also from newspaper sites, namely *The Metro*, *The Daily Express*, and *The Daily Mail*. The Portuguese subcorpus also resorts to YouTube comments, as well as to Portuguese newspaper sites, namely *O Público*, *Sol*, and *Observador*. Our software engineering team colleagues designed a set of bespoke extraction programmes and scraping tools, which had to be constantly customised so as to dodge the armoured online platforms and their ever-changing source codes (Henriques et al., 2019).

The primary aim of the project was to understand how user-generated content in social media expresses hate—i.e. prejudice and discrimination—against groups that are disadvantaged, be it in social, political, economic, legal, historical, physical, or symbolic terms.[2] So as to tackle the

[2] The phrase "disadvantaged groups" should be viewed from a variety of competing terms, such as "vulnerable groups", "oppressed groups", "protected groups", and "minority groups", each of which is bound to trigger a controversy of their own. See Chap. 2, Sect. 3.2.2.

variety and complexity of such an intricate object, the NETLANG team started by determining what exactly such groups might be and by arranging them into a programmatic table. A list of ten social variables (see e.g. Burgess, 2018) was formulated, including gender, ethnicity, age, nationality, and social class, as well as other identity-forging factors, namely sexual orientation, religion, gender identity, physical features (including disability issues), and behaviour issues (esp. drug abuse). Some of the variables were further divided into subtypes, in a hierarchical conceptual structure: for instance, the "gender" variable was broken down into different categories, namely "female physical appearance", "female sexuality", and "female intelligence", since women were deemed to be prevailing targets of gender discrimination, especially with regard to three defining features (their looks, sexual conduct, and intellectual capacity). Then, each social variable row on the table was matched with the sort of prejudice it typically produces: respectively, sexism, racism, ageism, nationalism, classism, homophobia, religious intolerance, transphobia, body shaming, and ableism.

The selection of the online texts was carried out by searching news articles and videos that were likely to arouse hateful responses. The way to do this was to keep track of daily news events around sensitive topics, capable of stirring hateful reactions, such as racial killings, the refugee crisis, euthanasia laws, social subsidies, transitioning, and domestic violence, among others. Thematic, documentary-type YouTube videos were also scrutinised, together with the "related content" list on the side, which supplies a sometimes long history of similar videos and comment trends.

Once the online texts were extracted, and turned into JSON files containing the entire comment threads existing at the time of extraction, they were automatically classified according to the table of social variables and types of prejudice. This was done by applying NetAC (Elias et al., 2021), a keyword analysis tool specifically designed for the NETLANG project, which uses a statistical framework to calculate the most frequent lexical items in the text, and automatically assign the corresponding prejudice label. The tool resorts to the list of keywords that the NETLANG team prepared for each of the social variables and corresponding types of prejudice, so as to provide a set of probable lexical cues signalling prejudiced discourse passages (on keyword shortcomings, see Scott, 2010;

MacAvaney et al., 2019, and Sect. 2 above). In the case of classist prejudice, for example, words deemed to prompt discriminatory content as regards social status were, among many others, Beggar, Boor, Bum, Bumpkin, Clodhopper, Hayseed, Hick, Hillbilly, Loser, Peasant, Pikey, Redneck, Scrounger, Trailer trash, Underdog, Vagrant, White trash, and Yokel. Similar keywords were devised for every other type of prejudice on the table. NetAC, therefore, automatically classifies each comment thread according to the presence of such keywords. It should be noted that many texts were classified under more than one type of prejudice (which, pardon the layman extrapolation, hints at the proverb that trouble never comes singly). For instance, racist comments were often found to be sexist as well. Alternatively, prejudiced comments frequently prompted an equally prejudiced reply by an interlocutor, but targeting a different social group.

A set of essential pre-processing operations followed the corpus assembling phase. Tokenisation, lemmatisation, and part-of-speech tagging were carried out, with a view to enabling linguistic annotation and computational analysis of the electronic texts. The particulars of online language, with its typical informality and speech-like nature (or "silent orality", as Soffer 2010 puts it), make it challenging for computational treatment. In the case of an online hate speech corpus, the issue of creativity, patent in deliberate misspellings and other strategies to avoid detection, presents further difficulties (see Dias & Pereira, this volume).

Constructing and preparing NETLANG's bilingual comparable corpus was the first step of the project. Analysing subsets of data extracted from the NETLANG corpus was the second step, which gave rise to seven of the chapters in the present collection. I will next introduce them, together with the rationale underlying the organisation of the book.

4 Book Layout: Parts and Chapters

The chapters in this book reflect the richness and variety of linguistic analyses which the NETLANG corpus understandably generates. So exciting are the online users' comment texts, despite their deep-seated negativity, so outrageous the dialogues they prompt and so intense the

intellectual and emotional interaction they express, that the present set of articles is but a sample of a very large territory to explore. The contributions that deal with other hate speech corpora, by researchers external to the NETLANG project, are no less stimulating in that they also look at various textual topics from rather different theoretical perspectives—and, it should be stressed, in a number of different languages.

Indeed, an important pointer to the polymorphous character of this book is the range of languages analysed: besides English and Portuguese, Danish, Lithuanian, Persian, Polish, and Slovenian data are examined. As a consequence, an array of geopolitical contexts for hate speech are discussed, especially featuring anti-refugee and anti-immigrant discourse, as is the case of Danish social media content against foreigners. Such contexts also involve stressful neighbouring relationships in countries like Poland and Ukraine, Iran and Afghanistan, and Israel and Palestine. In terms of analysis, the diversity of approaches is not only thematic but methodological: indeed, it resides not only in the various types of prejudice covered—from sexism to nationalism, racism, antisemitism, religious intolerance, ageism, and homo/transphobia—but also in the various analytical methods employed and theoretical frameworks used to support the linguistic analyses.

The three parts that compose this book are based on the different linguistic phenomena under focus, rather than on thematic organisation. If the latter were the case, the contributions could be grouped under the different types of prejudice and social variables covered, for instance sexism and gender issues, racism and ethnicity issues, and so on. Instead, the focus is on the angle and level of linguistic analysis, covering a variety of approaches and an array of overt to covert textual data, namely (i) structural and explicit elements, like syntactic and morphological patterns which recur throughout the texts, (ii) lexical and stylistic elements, aiming at the often implicit ways in which vocabulary choices and rhetorical devices signal the expression of hate, and (iii) interactional elements, focusing on the pragmatic relationships established in the online communicative exchanges. An important proviso is that this division is not rigid or clear-cut. As many of the chapters intersect, dealing with more than one textual feature, or even tackling multiple linguistic facets simultaneously and rather comprehensively, they could be viewed differently

and placed elsewhere. For the sake of simplification, then, let us consider the three broad Parts as mere indicators of the theoretical and methodological diversity of the volume.

The **Introduction** to the book is called "**Online Hate Speech: Object, Approaches, Issues**" and it is divided into two preliminary chapters, where I provide the general theoretical framework of the book and its design. The chapter you are reading now presents the collection and outlines the NETLANG project, from which it originates, as well as the external contributions to the study of online hate speech that integrate the volume. The second chapter is also of an introductory nature, as it is a definitional and programmatic reflection on the object under focus. What is hate speech, after all? An overview of the extant scholarship reveals a tendency towards imprecision, dismissiveness, and all-inclusiveness. "Hateful" is often found amidst a variety of related, but by no means synonymous, adjectives, such as aggressive, offensive, insulting, abusive, rude, toxic, unacceptable, and extreme. Yet, using aggressive language, no matter how violently, in situations of disagreement, dislike, or annoyance, ought to be distinguished from the intentional expression of prejudice towards (members of) disadvantaged groups, with likely harmful ideological agendas. By revisiting the classic communication framework involving senders, messages, channels, and receivers, and enlightening it with current linguistic insight, I put forth an annotation model of five factors, the positive co-occurrence of which signals hate speech. I then test the model against a range of examples from three subsets of the NETLANG corpus, classified under the sexism, racism, and ageism variables. Finally, I extend the analysis into other, linguistic, features of online hate speech. This preliminary, introductory discussion is therefore meant as a definitional layout of the object under focus in the book.

Part II of the volume, **"Structural Patterns in Hate Speech"**, covers four chapters, each dealing with a different grammatical structure or element that recurs in the corpus as a marker of hate: (a) regular expressions with a stable structure, (b) verbs used in the first person, (c) demonstrative determiners, and (d) compounds and syntactic patterns, all of which are found to reappear in sequences of prejudice expression. This set of studies illustrates treatments of hate speech that lie outside strict content word analysis and instead reside in form and grammar. Most automatic

tools designed to detect hate speech in large corpora have tended to concentrate on lexicon-based instruments targeting (negative) opinions end emotions. But the role of functional elements, such as determiners and conjunctions, in providing detection cues may prove very productive, supplying functionally embedded meanings. Besides, certain syntactic structures have been found to recur in hate speech corpora, which makes them promising research material.

The first chapter in Part II is by Idalete Dias and Filipa Pereira—respectively, the co-principal researcher of the NETLANG project and its full-time grantee. In "Improving NLP Techniques by Integrating Linguistic Input to Detect Hate Speech in CMC Corpora", they focus on automatic models that resort to Machine Learning (ML) and lexicon-based approaches in order to identify occurrences of hate speech in large corpora. They begin by describing details of the NETLANG corpus construction, namely of the Natural Language Processing (NLP) techniques used to obtain a tokenised, lemmatised, and part-of-speech tagged English-Portuguese corpus. Given the limitations of existing annotation tools, optimised for standard written production, instead of the highly informal and manipulated nature of CMC language, they discuss how their performance can be improved by integrating linguistic input. The mixed methods approach they put forth combines linguistic knowledge—including lexical, syntactic, and pragmatic input—with NLP techniques to trace fixed expressions conveying hate in user-generated content. Their aim is to analyse the behaviour of opinion markers that exhibit a certain degree of fixedness in the English subset of the NETLANG corpus as potential pointers to hateful (namely sexist and racist) content.

In Chap. 4, Ylva Biri, Laura Hekanaho, and Minna Palander-Collin (the latter a member of the NETLANG team), analyse the Sexism subset of the NETLANG corpus in English, with a view to identifying instances where YouTube commenters explicitly express their personal intention to carry out physical violence against individual targets. So as to do this, the Authors concentrate on aggression verbs in constructions with the first-person singular subject pronoun (e.g. *I punch*, *I kill*), which signal the speaker's assumption of an agentive position. The approach combines exploratory corpus-based investigation of patterns of 'I + aggression verb'

with an inductive approach, based on a qualitative, manual inspection of the comments in context. Taking advantage of the big data nature of the corpus, the Authors lay out a general taxonomy of first-person singular aggression online, expressing sexist and misogynistic language, including references to sexual assault. The taxonomy is broadly divided into "threats of physical aggression" (simple or conditional) and "expressions of mental aggression" (boulomaic and emotive), showing the variety of manifestations of violent and hostile verbal behaviour against women on the Internet.

Chapter 5, co-authored by Joana Aguiar and Pilar Barbosa, home members of the NETLANG team, sets out to explore the use of a single, yet manifold, grammatical element—demonstrative determiners—to convey the speaker's emotional involvement in, or detachment from, the subject under discussion, and to create solidarity effects between speaker and addressee. Deixis, a concept that usually concerns the speaker's spatial and temporal perception of certain objects or persons, is here understood in the Lakoffian sense of an emotional marker of closeness versus remoteness as regards the referent and/or the addressee. By focusing on the Portuguese subset of the NETLANG corpus, the Authors depart from the hypothesis that the three-way system of demonstratives in Portuguese may be used with pragmatic functions, namely to establish either proximal or distal evaluative meanings. After encoding all occurrences through a preliminary Sentiment Analysis approach, they examine the presence of two analytical variables: markers of exclamativity and proper nouns following the determiner. The results suggest that the interaction of the two variables may boost the affective potential of demonstrative determiners, making the combination a possible pointer to hate speech in computer-mediated discourse.

In Chap. 6, Eckhard Bick analyses a variety of language phenomena targeting immigrant and refugee minorities in Denmark, with a special emphasis on morphological processes of compound formation and on the role of syntax in generalisation and othering. Besides such specific grammar-based features, Bick also looks into more general prejudiced stereotypes and narratives, dehumanising metaphors, target-specific slurs, and the role of emojis as conveyors of sentiment and metaphor, making his chapter one of the more comprehensive accounts of hate-speech

language in the volume, escaping a clear categorisation. The analysis is based on a large social media corpus compiled under the auspices of a bilingual Danish-German project, XPEROHS, and annotated at the morphological, syntactic, and semantic levels. The chapter also sports quantitative, corpus-based methods meant to (i) enable qualitative examination of the data based on linguistic pattern searches, (ii) allow for the inspection of co-occurrences and relative frequencies, so as to identify typical features of target concepts, and (iii) devise sentiment ranking on the basis of word vector distances and machine-learned word embeddings.

Part III, **"Lexical and Rhetorical Strategies in the Expression of Hate Speech"**, deals with the way in which particular vocabulary choices and stylistic tactics manage to convey hateful content in strategically implicit ways. This covert expression of hate is achieved through indirect devices such as metaphor and irony, presupposed claims, and various allusions, not to mention mock politeness and humorous manoeuvres. Instead of being entertaining or solidarity-based, these creative moves may be meant to circumvent censorship, since the more manifest hatred is, the more probable to get reported. The four chapters in this Part explore different aspects of figurative language, rhetorical resources, specific epithets, and certain analogies which seem to be employed to conceal the otherwise more easily censored expression of prejudice.

In Chapter 7, Liisi Laineste and Władysław Chłopicki, both of whom were members of the NETLANG project, set out to inspect how figurative language, in particular complex metaphors, may be exploited to disguise the hateful import of religion-based comments in the NETLANG corpus. They start by discussing the propensity of the Internet to express negative emotions, and move on to ascertain how humour, together with figurative devices such as metaphor, irony and exaggeration, can mitigate such emotional negativity. By applying Deliberate Metaphor Theory to the analysis of a set of religious hate speech comments on YouTube, they focus on conventional and deliberate metaphor usage from the three-fold perspective of metaphor in mind, language, and communication. Then they concentrate on its crucial combination with humour, hypothesising that humorous metaphors play an important role in alleviating the hatefulness of the message and hence in avoiding detection. The analysis reveals the existence of a scale of social inappropriateness from the (less

inappropriate) complex, deliberate metaphors to the (more inappropriate) simpler, conventional ones, which suggests a continuum of metaphorically expressed hate as a function of context.

In Chap. 8, Vahid Parvaresh and Gemma Harvey examine the way in which hate is implicitly conveyed by means of rhetorical questions in Instagram comments targeted at Afghan people. The corpus comprises numerous comment threads totalling 700 individual comments published on the official Instagram account of Persian BBC between early 2019 and early 2021, i.e. prior to the taking over of Afghanistan by the Taliban. The findings indicate a ubiquitous use of rhetorical questions by Iranians in communicating meanings which in some way attack the dignity of Afghan people, express biased opinions about them or convey harmful stereotypes. Parvaresh and Harvey divide their analysis of occurrences of rhetorical questions into (i) those that evoke confirmatively positive responses to hateful illocutions, (ii) those that evoke confirmatively negative responses, (iii) those that do not allow any response but a negative one, and (iv) those that evoke a wider range of hateful responses. All such strategies seem to confirm the pervasive expression of prejudice and discrimination against a certain national and ethnic group, namely the Afghans, as well as the attempt to convey such hateful content indirectly and implicitly.

Chapter 9, co-authored by Matthew Bolton, Matthias J. Becker, Laura Ascone, and Karolina Placzynta, sets out to explore the role which so-called "enabling concepts" play in expediting and legitimating the expression of prejudiced ideas and concepts against Israel. By focusing on one such enabler, namely the Apartheid analogy, it offers a qualitative analysis of a set of more than 10,000 comments posted on 24 mainstream web news forums on the May 2021 Arab-Israeli conflict, where a range of anti-Semitic stereotypes, together with other analogies, surface in the vicinity of the Apartheid trope. These stereotypes are classed into four categories (*evil*, *guilt*, *power and instrumentalisation*, and *child murder*), whereas the analogies are grouped under the labels of *Nazism, colonialism, fascism*, and *terrorism*. The study springs from the *Decoding Antisemitism* project and is deemed to be pragmalinguistic, taking into account the immediate context of the comment thread and broader world knowledge. It departs from a general historical, political, and legal

overview of the meaning(s) of Apartheid and of how its purportedly widespread acceptability makes it an enabler of a range of anti-Semitic attacks, which would otherwise be more easily countered.

In Chap. 10, Lucyna Harmon offers an analysis of Twitter hateful content against the influx of Ukrainian refugees in Poland since the beginning of the Russian war on Ukraine. She starts by providing an overview of Polish-Ukrainian relations, including historic conflicts regarding territory boundaries, painful scars, like the Volhynian massacre, and ensuing resentments, which jeopardise the officially collaborative connection between the two neighbours, the epitome of which is the 2022 massive migration of 7.3 million Ukrainian people into Poland. Harmon then focuses on a set of Polish creative neologisms used on Twitter comments which constitute derogatory derivatives made on the basis of clippings and blends (such as *ukry*, *Banderowcy*, *upadlina*), used by Poles to designate, and denigrate, the Ukrainians. This word-formation analysis of hateful vocabulary, which includes metaphorical extensions, is accompanied by an examination of prejudiced narratives, like the purported Ukrainisation of Poland. Harmon's content analysis resorts to a classification according to Hart's (2010) CDA *topoi* list, including such categories as *burden*, *character*, and *danger*, among others. The overall analysis seems to confirm a biased portrayal of Ukrainians as usurpers and invaders, rather than victims, along the lines of the typical anti-refugee discourse that affects other nationalities seeking shelter.

Part IV, **"The Interactional Dimension of Hate Speech: Negotiating, Stance-Taking, Countering"**, looks at how the meanings of hate speech are dialogically constructed, on the basis of the interactional dynamics established between the various participants in the communicative situation. Even though most online platforms are asynchronous, the sort of conversation they enable is highly reactive, and it may be built as a result of, and at the same rate as, the very comments that are produced. The four chapters that constitute this final Part of the collection explore different facets of how social media commenters negotiate their participation, construct their contributions, react to their interlocutors' comments or to those directed at third parties, and assume certain personas and stances as the discussions, sometimes heatedly, unfold.

In Chap. 11, Rita Faria, a member of the NETLANG Project, sets out to examine how patterns of misogynistic hateful discourse emerge in the Portuguese newspaper section of the NETLANG corpus. She adopts a qualitative methodological approach which integrates different research strands, namely the guiding notion of stance-taking anchored in the Theory of Social Actors, elements of the Appraisal framework, and input from Impoliteness Studies, with a minor quantitative dimension offered to reinforce the qualitative analysis. Particular attention is paid to the discursive realisations of such linguistic devices as suffixation, possessivation, and collocations based on demonstrative determiners. By focusing on particular targets, namely women public figures (whether public office holders or well-known media figures), the analysis reveals how participants consistently take an aggressive, adversarial stance of misalignment and opposition, and voice openly discriminatory opinions about such women on the basis of their gender identity, thus reproducing a range of related sexist biases and stereotypes.

Chapter 12, by Jan Chovanec, also a member of the NETLANG project, departs from a discussion of such interrelated concepts as anti-social discourse, hate speech, aggressive speech, and conflict talk. By adopting a sociopragmatic conception of conflict talk as a multi-dimensional phenomenon with several key dimensions (structure, linguistic realisation, and meaning), it extends the notion of conflict by integrating its sociolinguistic indexicality, in terms of identity construction and status assertion. The analysis focuses on a range of discursive strategies that are employed to express conflict in a NETLANG subset on body shaming and physical impairments. It describes how commenters use conflicting representations, engage in extended conflictual discussions, and escalate the mutual conflict, while gradually shifting from idea-oriented to person-oriented strategies. The findings suggest that conflict can be exploited to delegitimise the other, while simultaneously strengthening the accord of the in-group, united in what Chovanec calls “harmony in hatred”.

In Chapter 13, Kristina Pahor de Maiti, Jasmin Franza, and Darja Fišer explore the role of gender in the production of online hate speech. More specifically, they examine gender differences regarding emotional expressiveness in what they call “socially unacceptable discourse” (SUD) online. By resorting to a corpus of English and Slovene Facebook

comments, they look at three levels of linguistic analysis—typographical, grammatical, and lexical—so as to assess the use of explicit markers of affect by males and females. Two research questions guide the analysis: whether men and women differ in the quantity of production, and whether women and men differ in their use of affective linguistic features in SUD comments. The results show statistically significant differences in the use of linguistic markers of affect between English-speaking or Slovene-speaking male and female commenters with regard to their hate speech production. As predicted, men are more likely to post shorter and more violent comments as opposed to offensive ones, while women tend to include more linguistic markers of affect in their comments on all levels of analysis.

Last but not least, Chap. 14, by Jūratė Ruzaitė, aims to study how participants in the comment section of a Lithuanian news portal resist and repel occurrences of LGBTQIA-targeted hate. Counterspeech, or bystander's intervention in attacks to third parties, which some advocate to be the key to combatting hate speech, is put under scrutiny, not only in terms of its incidence but especially of its linguistic construction. The chapter combines a Critical Discourse Analysis perspective with input from impoliteness theory and argumentation studies, especially in terms of *topoi* analysis. The quantitative results show that counterspeaking is scarce and more common only in the section of registered users. The qualitative analysis reveals that the argumentation used in counter-comments contains a high degree of hostility and often resembles that of homophobic comments, with the proviso that the target of the attacks is not a disadvantaged group. Only in a small number of comments does the argumentation aim at constructive dialogue. Even so, the cases analysed give promising evidence of the existence of alternative, resistance voices in a rather hegemonic prejudice-laden cyberspace.

One final note is in order in this introductory chapter. As is expectable in a book on hate speech, the following pages contain language and sequences of virtual dialogue which may be not only unpleasant but offensive, possibly disturbing, and even shocking at times. However, it is the linguist's onus to look into all things linguistic, including hateful and discriminatory communication. All the chapters in this collection analyse real, actual user-generated content, publicly posted in open online forums

with no registration required—which, according to the European Commission's guidance note (2021: 14), makes any expectations of privacy unreasonable. And even though the given (or purported, pseudonymous) identity of the commenters is not revealed, no change has been made to the actual phrasing of the comments. By extension they include all original insults, slurs, swearwords, taboo terms, etc., as well as all orthographic, grammatical, and typographical idiosyncrasies and infelicities of the users' texts. Of course, all verbatim occurrences of hateful language throughout the book must be viewed as instances of "mention" rather than "use" (e.g. Anderson & Lepore, 2013), that is, as "reported" hate speech, "mentioned" for the sake of academic inquiry.

Acknowledgement This work was sponsored by FCT (Foundation for Science and Technology, Portugal), under the auspices of the NETLANG project, ref. PTDC/LLT-LIN/29304/2017.

References

Anderson, L., & Lepore, E. (2013). Slurring words. *Noûs, 47*(1), 25–48.

Argueta, C., Calderon, F. H., & Chen, Y.-S. (2016). Multilingual emotion classifier using unsupervised pattern extraction from microblog data. *Intelligent Data Analysis, 20*, 1477–1502.

Assimakopoulos, S. (2020). Incitement to discriminatory hatred, illocution and perlocution. *Pragmatics and Society, 11*(2), 177–195.

Assimakopoulos, S., Muskat, R. V., Van Der Plas, L., & Gatt, A. (2020). *Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis*. arXiv preprint arXiv:2008.06222.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA).

Badlani, R., Asnani, N., & Rai, M. (2019). Disambiguating sentiment: An ensemble of humour, sarcasm, and hate speech features for sentiment classification. *W-NUT, 2019*, 337–345.

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology, 24*(3), 233–239.

Baumgarten, N., Bick, E., Geyer, K., Iversen, D. A., Kleene, A., Lindø, A. V., Neitsch, J., Niebuhr, O., Nielsen, R., & Petersen, E. N. (2019). Towards balance and boundaries in public discourse: Expressing and perceiving online hate speech (XPEROHS). In J. Mey, J. A. Holsting, & C. Johannessen (Eds.), *RASK – International Journal of Language and Communication* (Vol. 50, pp. 87–108). University of Southern Denmark.

Beliaeva, N. (2022). Is play on words fair play or dirty play? On ill-meaning use of morphological blending. In N. Knoblock (Ed.), *The Grammar of Hate* (pp. 177–196). Cambridge University Press.

Benesch, S. (2014). Defining and diminishing hate speech. *State of the World's Minorities and Indigenous Peoples*, 18–25.

Bianchi, R. (2022). 'Kill the invaders': Imperative verbs and their grammatical patients in Tarrant's the great replacement. In N. Knoblock (Ed.), *The Grammar of Hate* (pp. 222–240). Cambridge University Press.

Brindle, A. (2016). *The Language of Hate: A Corpus Linguistic Analysis of White Supremacist Language*. Routledge.

Burgess, R. (2018). *Key Variables in Social Investigation*. Routledge.

Burke, S., Diba, P., & Antonopoulos, G. A. (2020). 'You sick, twisted messes': The use of argument and reasoning in Islamophobic and anti-Semitic discussions on Facebook. *Discourse and Society, 31*(4), 374–389.

Butler, J. (1997). *Excitable Speech: A Politics of the Performative*. Routledge.

Carney, T. (2014). Being (im)polite: A forensic linguistic approach to interpreting a hate speech case. *Language Matters, 45*(3), 325–341.

Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic Journal of Communication, 23*(1), 46–65.

Carr, C., Robinson, M., & Palmer, A. (2020). Improving hate speech detection precision through an impoliteness annotation scheme. In *94th Annual Meeting of the Linguistic Society of America*, New Orleans.

Culpeper, J. (2021). Impoliteness and hate speech: Compare and contrast. *Journal of Pragmatics, 179*, 4–11.

Culpeper, J., Iganski, P., & Sweiry, A. (2017). Linguistic impoliteness and religiously aggravated hate crime in England and Wales. *Journal of Language Aggression and Conflict, 5*(1), 1–29.

Dadvar, M., Jong, F., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the 12th Dutch-Belgian information retrieval workshop* (pp. 23–25).

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media, 11*, 512–515.

Dinakar, K., Reichart, R., & Lieberman, H. (2011). Modeling the detection of textual cyberbullying. *Soc. Mobile Web, 11*, 02.

Domínguez-Armas, A., Soria-Ruiz, A., & Lewiński, M. (2023). Provocative insinuations as hate speech: Argumentative functions of mentioning ethnicity in headlines. *Topoi, 42*, 1–13.

Đorđević, J. P. (2019). The sociocognitive dimension of hate speech in readers' comments on Serbian news websites. *Discourse, Context and Media., 33*(2020), 1–9.

Ehrlich, S. (2001). *Representing Rape: Language and Sexual Consent*. Psychology Press.

Ehrlich, S. (2014). Language, gender, and sexual violence. In S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The Handbook of Language, Gender, and Sexuality* (2nd ed., pp. 452–470). Wiley.

Elias, C., Gonçalves, J., Araújo, M., Pinheiro, P., Araújo, C., & Henriques, P. (2021). NetAC, an automatic classifier of online hate speech comments. In A. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. M. Ramalho Correia (Eds.), *Trends and Applications in Information Systems and Technologies* (pp. 494–505). Springer.

Ermida, I. (2014). A beached whale posing in lingerie: Conflict talk, disagreement and impoliteness in online newspaper commentary. *Diacrítica, 27*(1), 95–130.

Ermida, I. (2018). 'Get the snip – And a job!' Disagreement, impoliteness, and conflicting identities on the internet. *Token: A Journal of English Linguistics, 6*, 205–247.

Ermida, I., Pereira, F. & Dias, I. (2023). Social media mining for hate speech detection: Opinion and emotion conflict in adversative constructions. Forthcoming.

Esposito, E., & Zollo, S. A. (2021). "How dare you call her a pig! I know several pigs who would be upset if they knew": A multimodal critical discursive approach to online misogyny against UK MPs on YouTube. *Journal of Language Aggression and Conflict, 9*(1), 47–75.

European Commission. (2021). *Ethics and Data Protection*. Guidance note by DG Research and Innovation. Retrieved from https://ec.europa.eu

Fersini, E., Nozza, D., & Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *EVALITA – Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop*, pp. 59–66.

Gehl, R. W. (2016). Power/freedom on the dark web: A digital ethnography of the dark web social network. *New Media & Society, 18*(7), 1219–1235.

Gelber, K. (2017). Hate speech definitions & empirical evidence. *Constitutional Commentary, 32*, 619–629.

Geyer, K., Bick, E., & Kleene, A. (2022). 'I am no racist but...': A corpus-based analysis of xenophobic hate speech constructions in Danish and German social media discourse. In E. Knoblock (Ed.), *The Grammar of Hate* (pp. 241–261). Cambridge University Press.

Gilbert, O., Pérez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online* (pp. 11–20). Association for Computational Linguistics.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*, 215–230.

Godioli, A., Young, J., & Fiori, B. M. (2022). Laughing matters: Humor, free speech and hate speech at the European court of human rights. *International Journal for the Semiotics of Law/Revue internationale de Sémiotique juridique, 35*, 1–25.

Hardaker, C., & McGlashan, M. (2016). 'Real men don't hate women': Twitter rape threats and group identity. *Journal of Pragmatics, 91*, 80–93.

Hart, C. (2010). *Critical Discourse Analysis and Cognitive Science: New Perspectives on Immigration Discourse*. Palgrave Macmillan.

Hedger, J. A. (2013). Meaning and racial slurs: Derogatory epithets and the semantics/pragmatics interface. *Language & Communication, 33*(3), 205–213.

Henriques, P., Araújo, C., Ermida, I., & Dias, I. (2019). Scraping news sites and social networks for prejudice term analysis. In H. Weghorn & L. Rodrigues (Eds.), *Proceedings of the 16th International Conference on Applied Computing 2019* (pp. 179–189).

Hornsby, J. (1995). Speech acts and pornography. In S. Dwyer (Ed.), *The Problem of Pornography*. Wadsworth Publishing Company, Springer.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*, 216–225.

Kazienko, P., Bielaniewicz, J., Gruza, M., Kanclerz, K., Karanowski, K., Miłkowski, P., & Kocoń, J. (2023). Human-centred neural reasoning for subjective content processing: Hate speech, emotions, and humor. *Information Fusion, 94*, 43–65.

Kienpointner, M. (2018). Impoliteness online: Hate speech in online interactions. *Internet Pragmatics, 1*(2), 329–351.

Knoblock, N. (Ed.). (2022). *The Grammar of Hate: Morphosyntactic Features of Hateful, Aggressive, and Dehumanizing Discourse*. Cambridge University Press.

Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet, 13*(4), 502–521.

Korecky-Kröll, K., & Dressler, W. (2022). Expressive German adjective and noun compounds in aggressive discourse. In N. Knoblock (Ed.), *The Grammar of Hate* (p. 197). Cambridge University Press.

Lange, P. G. (2014). Commenting on YouTube rants: Perceptions of inappropriateness or civic engagement? *Journal of Pragmatics, 73*, 53–65.

Langton, R. (1993). Speech acts and unspeakable acts. *Philosophy and Public Affairs, 22*(4), 293–330.

Langton, R. (2018). The authority of hate speech. *Oxford Studies in Philosophy of Law, 3*, 123–152.

Lederer, L. J., & Delgado, R. (Eds.). (1995). *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*. Hill & Wang.

Leskova, A. (2016). *"Black Humor" in Modern Europe: Freedom of Speech v. Racist Hate Speech. Or Where is the Line for Racist Humor?* Doctoral dissertation, University of Sevilla.

Lewis, M. (2012). *A Cognitive Linguistics Overview of Offense and Hate Speech*. Available at SSRN 2205178.

Lind, M., & Nübling, D. (2022). The neutering neuter. The discursive use of German grammatical gender in dehumanisation. In N. Knoblock (Ed.), *The Grammar of Hate* (pp. 118–139). Cambridge University Press.

Liu, S., & Forss, T. (2015). New classification models for detecting hate and violence web content. In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K'15)* (Vol. 1, pp. 487–495). IEEE.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Proceedings of 22nd International Conference on Text, Speech, and Dialogue, TSD 2019* (pp. 103–114). Springer.

Lorenzo-Dus, N., Blitvich, P. G.-C., & Bou-Franch, P. (2011). On-line polylogues and impoliteness: The case of postings sent in response to the Obama Reggaeton YouTube video. *Journal of Pragmatics, 43*, 2578–2593.

MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS One, 14*(8), e0221152.

Macdonald, S., & Lorenzo-Dus, N. (2020). Intentional and performative persuasion: The linguistic basis for criminalizing the (direct and indirect) encouragement of terrorism. *Criminal Law Forum, 31*(4), 473–512.

MacKinnon, C. A. (1993). *Only Words*. Harvard University Press.

Matsuda, M. J., Lawrence, C. L., Delgado, R., & Crenshaw, K. W. (1993). *Words that Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*. Westview Press.

Mattiello, E. (2022). Language aggression in English slang: The case of the-o suffix. In N. Knoblock (Ed.), *The Grammar of Hate* (pp. 34–58). Cambridge University Press.

Menon, P. (2022). *Laughter is the Best Poison: Antagonistic Humor as the Handmaiden of Hate Speech*. University of Michigan – Ann Arbor.

Musolff, A. (2017). Dehumanizing metaphors in UK immigrant debates in press and online media. *Journal of Language Aggression and Conflict, 3*(1), 41–56.

Nagle, J. C. (2009). The idea of pollution. *UC Davis Law Review, 43*(1), 1–78.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, 145–153.

Nunberg, G. (2018). The social life of slurs. In D. Fogal, D. Harris, & M. Moss (Eds.), *New work on Speech Acts* (pp. 237–295). Oxford University Press.

O'Driscoll, J. (2020). *Offensive Language: Taboo, Offence and Social Control*. Bloomsbury.

Ohlson, L. F. (2022). The power of a pronoun. In N. Knoblock (Ed.), *The Grammar of Hate* (pp. 161–176). Cambridge University Press.

Özarslan, Z. (2014). Introducing two new terms into the literature of hate speech, "hate discourse" and "hate speech act": Application of speech act theory to hate speech studies in the era of web 2.0. *Galatasaray Üniversitesi İletişim Dergisi, 20*, 53–75.

Pettersson, K., & Sakki, I. (2023). 'You truly are the worst kind of racist!': Argumentation and polarization in online discussions around gender and radical-right populism. *British Journal of Social Psychology, 62*(1), 119–135.

Pražmo, E. (2020). Foids are worse than animals. A cognitive linguistics analysis of dehumanizing metaphors in online discourse. *Topics in Linguistics, 21*(2), 16–27.

Raffone, A. (2022). "Her leg didn't fully load in": A digitally-mediated social-semiotic critical discourse analysis of disability hate speech on TikTok. *International Journal of Language Studies, 16*(4), 17–42.

Raj, S. M., & Usman, A. (2021). The use of mental spaces in the conceptualization of hate speech. *GPH - International Journal of Social Science and Humanities Research, 4*(05), 12–21.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *The 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media of the Valencia Association for Computational Linguistics*, 1–10.

Schwartzman, L. H. (2002). Hate speech, illocution, and social context: A critique of Judith Butler. *Journal of Social Philosophy, 33*(3), 421–441.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails…. In M. Bondi & M. Scott (Eds.), *Keyness in Texts* (pp. 43–58). John Benjamins.

Sharifi, M., Ansari, N., & Asadollahzadeh, M. (2017). A critical discourse analytic approach to the discursive construction of Islam in Western talk shows: The case of CNN talk shows. *International Communication Gazette, 79*(1), 45–63.

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *ICWSM*, 687–690.

Sirulhaq, A., Yuwono, U., & Muta'ali, A. (2023). Why do we need a sociocognitive-CDA in hate speech studies? A corpus-based systematic review. *Discourse & Society*. https://doi.org/10.1177/0957926522112659

Soffer, O. (2010). "Silent orality": Toward a conceptualization of the digital oral features in CMC and SMS texts. *Communication Theory, 20*(4), 387–404.

Stokoe, E., & Edwards, D. (2007). 'Black this, black that': Racial insults and reported speech in neighbour complaints and police interrogations. *Discourse & Society, 18*, 337–372.

Tarasova, E., & Fajardo, J. A. S. (2022). Adj+ie/y nominalizations in contemporary English: From diminution to pejoration. In N. Noblock (Ed.), *The Grammar of Hate* (pp. 59–73). Cambridge University Press.

Trindade, L. V. P. (2020). Disparagement humour and gendered racism on social media in Brazil. *Ethnic and Racial Studies, 43*(15), 2766–2784.

Tsesis, A. (2009). Dignity and speech: The regulation of hate speech in a democracy. *Wake Forest Law Review, 44*, 497–532.

Van Dijk, T. A. (1987). *Racism and the Press*. Routledge.

Van Dijk, T.A. (2005). *Racism and discourse in Spain and Latin America*. Benjamins.

Van Dijk, T. A. (2021). *Antiracist Discourse. Theory and History of a Macromovement*. Cambridge University Press.

Vasilaki, M. (2014). Name-calling in Greek YouTube comments. In C. Pérez-Arredondo, M. Calderón-López, H. Hidalgo-Avilés, & D. Pask-Hughes (Eds.), *Papers from the 9th Lancaster University postgraduate conference in Linguistics & Language Teaching* (pp. 90–110). Lancaster University.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media of the Association for Computational Linguistics*, 19–26.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pp 88–93.

Weston, D. A. (2022). When does speech perform regulable action? A critique of speech act theory's application to free speech regulation. *International Journal of Language & Law (JLL), 11*, 78–97.

Williamson, T. (2009). Reference, inference and the semantics of pejoratives. In J. In Almog & P. Leonardi (Eds.), *The philosophy of David Kaplan* (pp. 137–158). Oxford University Press.

Woods, F. A., & Ruscher, J. B. (2021). Viral sticks, virtual stones: Addressing anonymous hate speech online. *Patterns of Prejudice, 55*(3), 265–289.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). *Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)*. arXiv preprint arXiv:1903.08983.