Studii de lingvistică 12, nr.2, 2022, 263 - 296

Evaluating Hate speech: A sociolinguistic and bio signal experiment

Fabienne Baider[1]
Christiana Anaxagorou[2]

Abstract: In the last ten years there have been many studies investigating the surge in online hate speech. Most such research is found in legal studies, discourse studies, and computer science studies. However, to date there has been very little *interdisciplinary experimental* research examining hate speech. Our study aims to fill this gap and to determine the efficacy of working within an interdisciplinary approach to study reactions to hate speech. We investigate reactions to *specific hate speech experiences* (e.g., for sexual orientation, racial profile, etc.) and then measure the reactions to these "stimuli" following a triangulated methodology (questionnaires, interviews and bio signal experiments). This paper discusses the results of such procedures after having given a comprehensive overview of hate speech studies.

Key words: Hate speech, sociolinguistic study, psychological experiment, bio signals, experimental linguistics.

1. Introduction[3]

Many studies over the last ten years have been devoted to the analysis of the surge in hate speech. Most such research is found in legal studies (Brown 2017, Enarsson and Lindgren 2019, Bayer and Bard 2020, Paz *et al.* 2020), psychological studies (Stephan *et al.* 1999), discourse studies (Erjavec and Kovačič 2012, Gagliardone *et al.* 2015, Assimakopoulos, Baider and Millar 2017, Rasaq *et al.* 2017), and computer science studies (Davani *et al.* 2021, Srba *et al.* 2021, Wachs *et al.* 2021, Švec *et al.* 2018, Poria *et al.* 2017). However, to date there has been very little *experimental* linguistic research examining hate speech, although there has been some recent psychological

[1] University of Cyprus; helenafab@yahoo.fr.
[2] University of Cyprus; anaxagorou.christiana@ucy.ac.cy.
[3] Fabienne Baider organized the project, wrote the article and did the sociolinguistic section. Christiana Anaxagorou analyzed the psychological experimental data.

experimental research (Neitsch and Niehbur 2020, Baumgarten *et al.* 2019). Our research project aims to fill this gap and has a twofold focus: first, to show how a sociolinguistic survey can shed light on the impact of hate speech; second, to examine whether psychological input, i.e., experimental analysis, can help in evaluating the impact of hate speech. We began by using a sociolinguistic methodology that included interviews (40 participants) and online questionnaires (around 100), in order to understand the *social* and *discursive* reactions to hate speech. Based on these findings we then developed a psychological experiment, which involved testing a small dataset of images and texts for their positive or negative effect on 40 participants. The testing procedure, which used the Open Sesame software (https://osdoc.cogsci.nl/), was carried out in four stages, in each of which participants viewed one short text and one meme related to homophobia, racism, migration, and sexism. This paper discusses the procedure and the results after having presented a comprehensive picture of hate speech studies carried out from an experimental perspective.

2. Context of the Study

2.1. Hate speech studies and experimental studies

Annual reports describing online hate speech within the European Union are deeply worrying. Recognizing that the concept of "hate speech" does not yet have a universal definition (Gagliardone *et al.* 2015, Fortuna and Nunes 2018, Baider 2020), for our experiment protocol we used the definition in Article 20 of the *International Covenant on Civil and Political Rights* or ICCPR. Here, hate speech is described as an "advocacy of discriminatory hatred which constitutes incitement to hostility, discrimination or violence". The 2019 ECRI report observed a sharp increase in both racist insults, which have become increasingly common, and xenophobic hate speech. As early as 2002, a study by Herring (2002) found that the Internet was indeed facilitating the global spread of hate.

If there are a plethora of linguistic studies focused on hate speech, as mentioned earlier (Erjavec and Kovačič 2012, Gagliardone *et al.* 2015, Rasaq *et al.* 2017, Baider 2020), very few researchers have used experiments to study hate speech and its impact.

Indeed, most research outside discourse analysis are focused on three domains: Hate Speech Detection (domain 1), Hate speech diffusion (domain 2) and Psychological profiles of hate speech producers (domain 3) (Masud *et al.* 2022).

Automatic hate speech detection is by far the most common type of research using technology and experiments. Some research focuses on keywords, linguistic characteristics, and textual content

(Burnap and Williams 2016, Waseem and Hovy 2016, Zhang and Luo. 2019), but more complex approaches consider the context (Cheng *et al.* 2020) and the use of images (Das *et al.* 2020, Kiela *et al.* 2020).

The propagation of hate speech and the dynamics involved in this process are the main topics researched in the Hate Speech Diffusion domain. Research attempts to understand the drivers of hatred by observing and analysing the ways online platforms facilitate the spreading of hateful content enable us to predict hateful replies (Zampieri *et al.* 2019).

The objective of domain 3 studies was the hate spreaders' psychology and the parameters involved in the judgements of hate speech. Some of this research attempts to model the archetype of "hate spreaders" using personality traits (Fischer *et al.* 2018), human values and social orientation (Karlekar and Bansal 2018), confirmation and / or social bias such as the use of stereotypes (Sap *et al.* 2020). Finally, and in the same field of research, judgments of offensiveness research have pointed to the complexity and contextuality of such judgements (Cowan and Hodge 1996, Guillén-Nieto 2020 and 2022, Almagro *et al.* 2022).

This earlier research is important since it highlights the difficulty in generalizing results given the contextual dimension of response and judgements. For example, ethnic speech was rated more offensive than gender- or gay-targeted speech in some contexts (Cowan and Hodge 1996), while the gender and ethnicity of the raters had an impact "on the effects of the experimental variables, as well as showing main effects" (Cowan and Hodge 1996: 355).

2.2. Bio signals and hate speech

We can note a few interdisciplinary experimental research studies related to the perception of hate speech involving linguistics and other disciplines. Baumgartner *et al.* (2019) have worked with explicit perceiver ratings (made by participants clicking onto scales), and Neitsch and Niehbur (2020) have explored innovative 2D rating spaces. Bio- signals have been also recently investigated (Neitsch and Niehbur 2020) and our study is similar to this latest study. Bio signals are a direct manifestation of the (sympathetic) nervous system, and three bio-signals were monitored in this pilot study: Heart Rate (HR), Breathing (BR), and Skin-Conductance Response (SCR). Each bio signal is determined and measured with specific tools. Such signals are less influenced by participants' conscious reflection and by any possible efforts to correct or change their behavior.

Bio-signals have rarely been used in hate speech studies; however they are interesting since the data collected is spontaneous, quickly collected, and more reliable insofar as the data are "less prone to interpretation biases" (Neitsch and Niehbur 2020: 710).

In Niehbur and Neitsch's 2020 study, the bio signals selected were correlated with mental stress and emotional arousal, and as such are considered reliable data to detect reactions to hate speech. In their study the bio-signals were used to determine "whether bio-signals mirror *explicit ratings* and are hence a suitable alternative in assessing the perception of hate speech" (1, our italics). Two hypotheses drove the study:

- *Explicit* perceiver ratings of overt hate speech should show an increase in all the HR/BR/SCR values for hate speech, whilst listening to covert hate speech (such as irony or rhetorical questions) would not show such an increase;
- *Spoken* stimuli would be perceived with more intense bio signals than written hate-speech stimuli.

Both hypotheses were confirmed in their experiment.

2.3. HOPE program: objectives and methodology

The work presented and discussed in the present article is part of the H.O.P.E. research program[4] focused on hate speech and counter narratives and funded by the University of Cyprus (2019-2021). Here we present our research focused on hate speech, which we studied according to the basic types of triangulation described by Denzin (1978):

- *Data triangulation:* involves time, space, and people. We adapted a questionnaire and an interview already used in the EU project C.O.N.T.A.C.T. (2015- 2017)[5], which allowed us to assess the strength of the results obtained in 2016 and within the same Greek Cypriot population.
- *Investigator triangulation:* involves multiple researchers in an investigation. We had three researchers involved in the interviews, while the experiment was also prepared by the whole team, with the stimuli chosen by each team member.
- *Methodological triangulation:* involves using more than one method to gather data. We used a sociolinguistic approach with interviews and questionnaires, and also adopted a psychological approach with the measurement of bio signals such as pulse and response time.

The aim of the experiment was to assess the impact of hate speech. To this end, we tested two questions identified in our previous

[4] 'Hate On line, Promoting Empathy'.
[5] https://portal.findresearcher.sdu.dk/en/publications/hate-speech-in-the-eu-and-the-contact-project.

research data Do hate speech texts trigger a more intense response in participants than hate speech images? Do texts convey hate more effectively than images?[6]

The first part of the research adopted a sociolinguistic approach and comprised online questionnaires and face-to-face interviews, the results of which would help determine the stimuli for our experiment. The same results were also be compared with the reactions obtained with bio signals. Thus, we first examined the findings of the online survey, which helped us determine the content for our semi-structured interviews. For example, the online questionnaire revealed that sexism was as widespread as racism and homophobia – results that informed both the interviews and the psychological experiment.

The second part of the research was based on a psychological framework inspired by Niehbur and Neitsch's (2020) experiments. We used memes and tweets illustrating data identified in the survey and interviews in order to highlight reactions to specific stimuli of hate speech. For example, we found that homophobia was more acceptable in the online survey and during the interviews; we used some memes resembling the arguments used to justify homophobia, and also added sexist memes and tweets because of the online survey findings.

This experiment, in conjunction with our sociolinguistic approach, aimed to offer insight into the impact of different categories of online and offline hate speech.

3. The Online Survey

The GDPR rules and bioethics guidelines were ensured by having only adult participants (above 18), who signed a consent form that described the experiment, its purpose, their rights to stop at any time and their right to know the results. All data are treated as confidential, i.e., participants chose their own coding (mixture of numbers and letters). We did, however, know the gender, level of education and age of the participants.

3.1. The online questionnaire: perception and experience of hate speech

The sociolinguistic study consisted of 120 questionnaires and 40 interviews. We had to discard 19 questionnaires, as these were incomplete. The analysis below is based on the answers of 95 respondents.

[6] We also investigated the impact of counter speech, but do not include those results. Counter- speech means arguments presented to deconstruct, respond or debunk hate speech.

The questionnaire contained 14 items divided into four sections, structured as follows:

Sections 1 and 2 include six statements used in a previous study (Baider 2019), some of which are clearly hate speech (racism and islamophobia). The statements were found in online comments on articles about gay pride or immigration. The participants had to record their opinion of the statements on a scale ranging from acceptable to unacceptable:

1	2	3	4	5
Acceptable	Somewhat acceptable	Less acceptable	Not acceptable	No answer

In section 3 participants rated their own experience of hate speech, personally or as a witness to such speech, and about their reactions and the reactions of others (7-12).
In section 4 participants shared their thoughts about the questionnaire, the hate speech phenomenon (13), and some personal details (14).

3.2. Results and analysis

As was the case in 2016, all overt hate speech statements were overwhelmingly rejected by participants (95% and 97%, respectively; see Figure 1):

> Statement 1: *Αυτή είναι η ράτσα που θα πρέπει να αφανιστεί απο τον πλανήτη. Ουστ* 'This is the race that needs to be annihilated from this planet. Out'
>
> Statement 2: *Μόνο νεκρός μουσουλμάνος καλός* 'The only good Muslim is a dead Muslim'

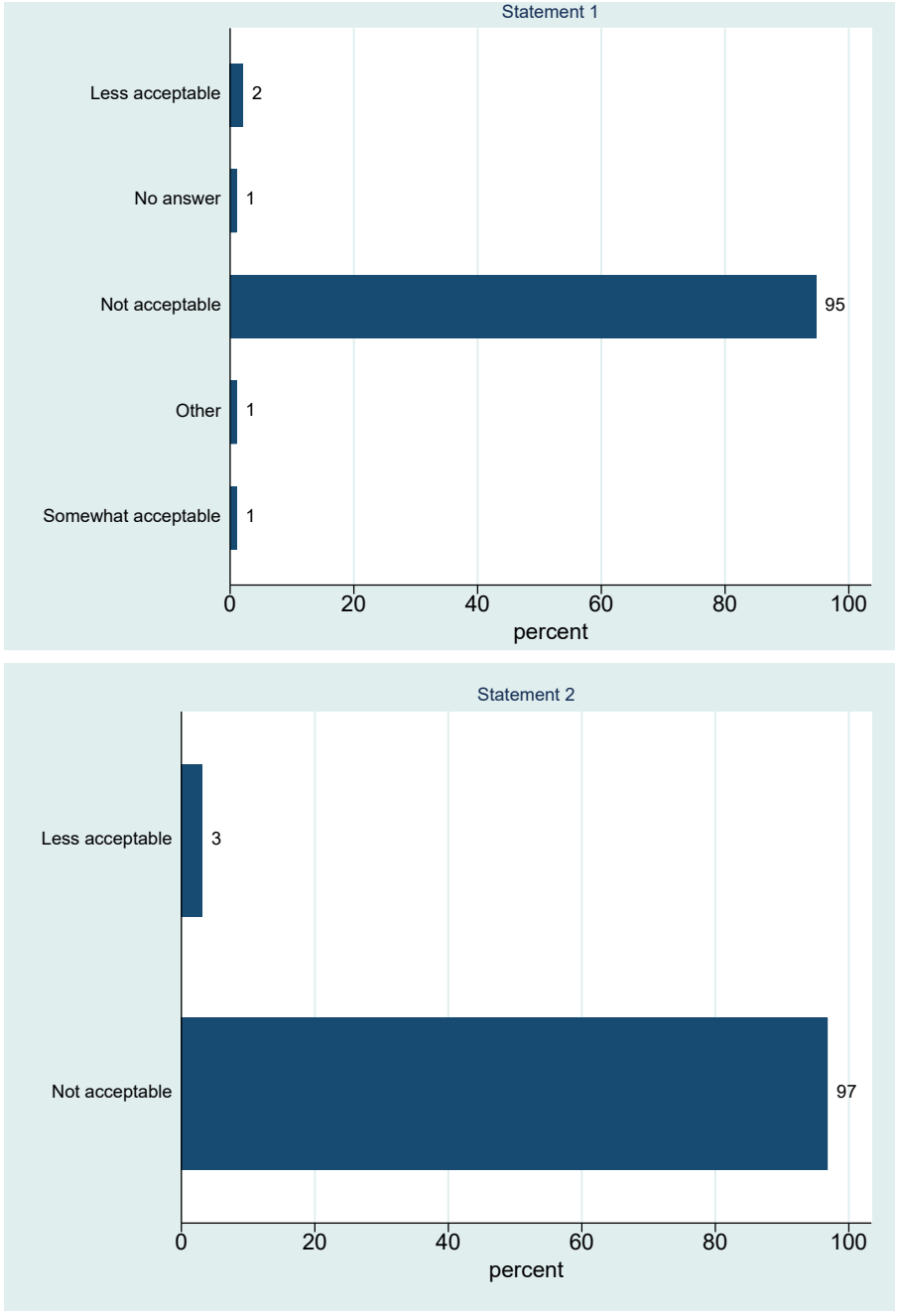
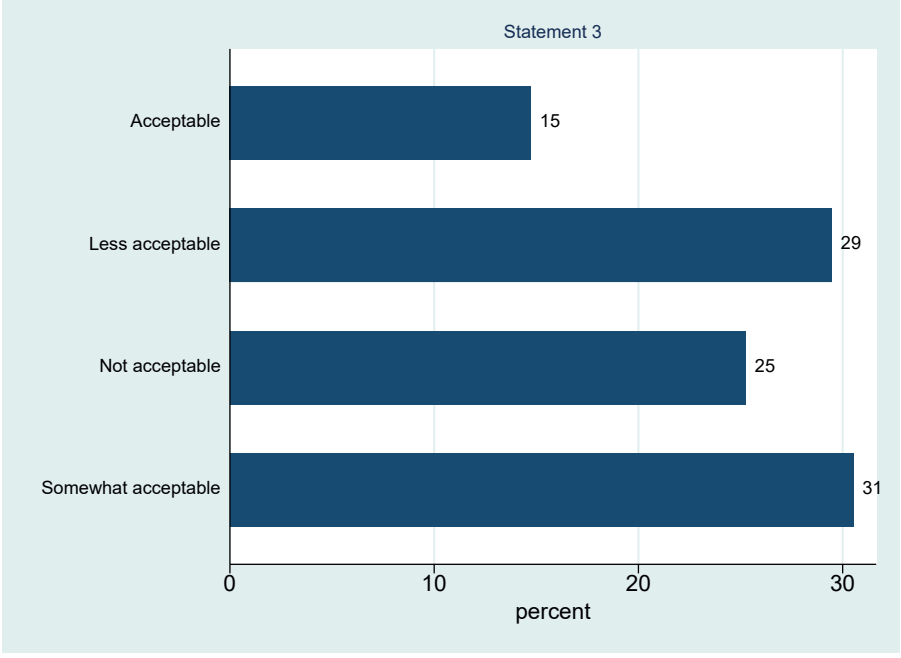


Figure 1: Acceptability of Hate speech

Statement 3, which refers to foreigners doing jobs Cypriots do not want to do, is not widely accepted (only 15% of the participants judge this sentence acceptable. This is rather surprising given the high number of female domestic workers from the Philippines and Sri Lanka who work in often very difficult circumstances, e.g., long hours, many more than they are contracted for (Hadjigeorgiou 2020)[7]. An equally high number of Eastern Europeans and Asian workers work in construction, often in very high temperatures and unsafe conditions for a monthly salary of less than 800 euros[8].

Statement 4 is a homophobic statement but without the context: *Προς το παρόν σας ανεχόμαστε, απλά μη παίζετε με την νοημοσύνη μας και μην τολμήσετε να δοκιμάσετε τα όρια της υπομονής μας. Άτε γιατί αρκετά σας ανεχτήκαμε.* 'At the moment we tolerate you, but don't think we are fools and don't dare try our patience. We have tolerated you enough'. It could also be a racist rant, and was rated unacceptable by 93% of participants.



[7] N. Hadjigeorgiou (2020), Report on the status of foreign domestic workers in Cyprus https://equineteurope.org/wp-content/uploads/2021/01/Cyprus-Domestic-Workers.pdf

[8] https://www.eurofound.europa.eu/publications/report/2007/employment-and-working-conditions-of-migrant-workers-cyprus; the figures are old since very few surveys are carried out to update the data.

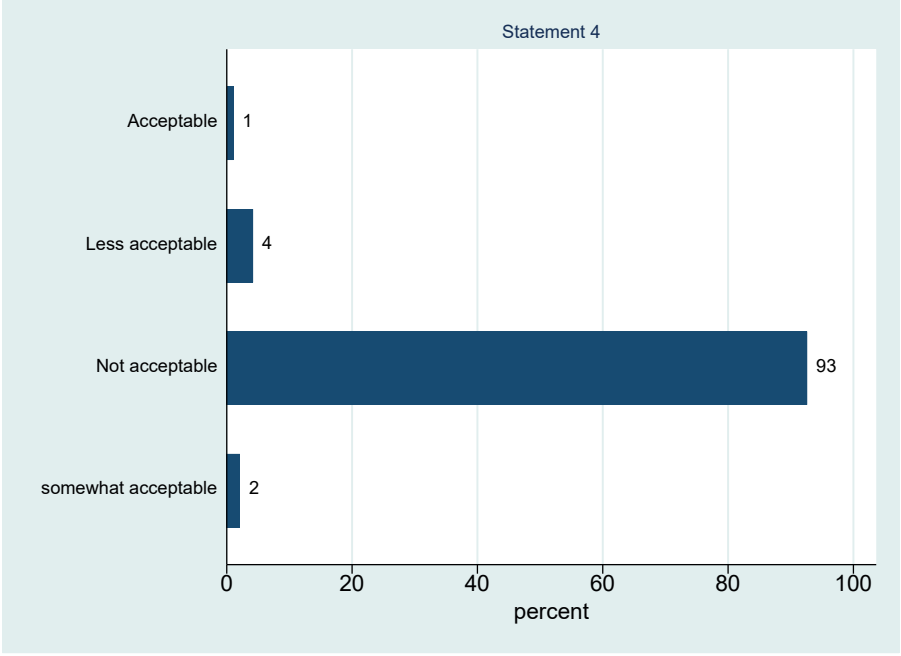


Figure 2: Acceptability of mixed statements

Section 2 of the questionnaire is devoted to homophobia, and in Figure 3 below, we see the mixed acceptability of gay rights in Cyprus.

Statement 5 refers to the legalization of the civil union for same sex couples in December 2015: *Η προώθηση της ομοφυλοφιλίας και άλλων μορφών αποκλίνουσας σεξουαλικής συμπεριφοράς, όπως έχουμε δει από την τρέχουσα μάχη για τον 'γάμο μεταξύ ομοφυλόφιλων', έχει σχεδιαστεί για να υπονομεύσει τόσο του γάμο και την οικογένεια όσο τους φυσικούς νομούς και την παραδοσιακή ηθική.* 'The promotion of homosexuality and other deviant sexual behavior, as we have seen in the current fight for "gay marriage" has been designed to undermine marriage and family as well as natural norms and traditional morality'. The statement equates homosexuality with deviant behavior and considers the fight to legalize same-sex civil unions as a fight against traditional values such as family. 73% found the statement unacceptable. This leads us to anticipate more negative reactions to racist stances than to homophobic statements when we test for bio signal reactions.

Similarly, only 74% found statement 6 acceptable: *Η σεξουαλικότητα του ανθρώπου δεν διδάσκεται, δεν επιβάλλεται, δεν ελέγχεται, δεν καθοδηγείται, δεν ακυρώνεται, δεν θεραπεύεται* 'A person's sexuality cannot be taught, imposed, controlled, guided, cancelled, cured'.

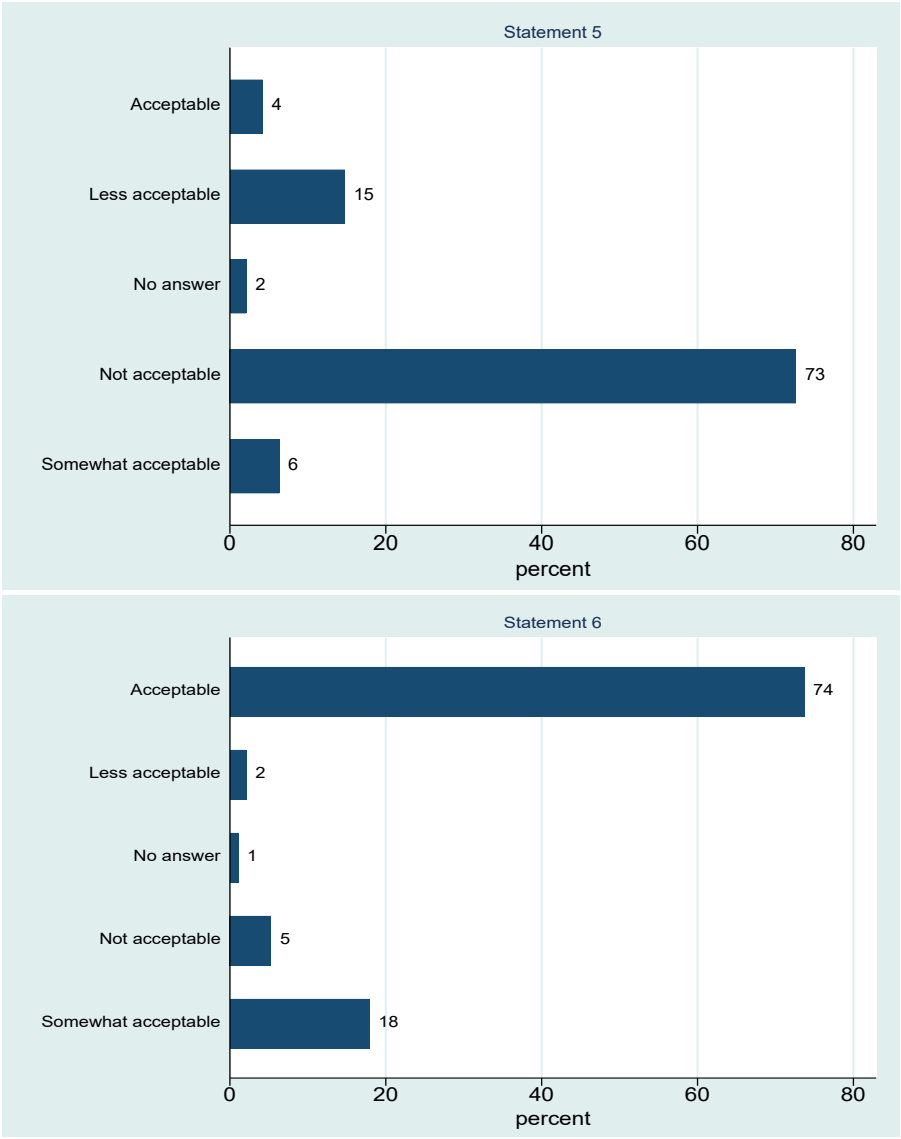


Figure 3: Acceptability of statements pro or against gay rights

In terms of participants' personal experience of hate speech – *Υπήρξατε ποτέ θύμα προσβολών ή απειλών λόγω του/της δικού σας:* 'Have you ever been a target of insults or threats because of your:', the majority responded negatively. The one exception related to gender (36%) indicated that including sexism in our experiment would be justified. This is important because most hate speech laws, such as the 2008 European Council decision, exclude gender.

Our participants reported that most hate speech incidents occurred on the street (28%), at work (22%), on public transport (19%) and online (14%); the Internet is not the first place where people are the most harassed (Herring *et al.* 2002, Brown 2017).

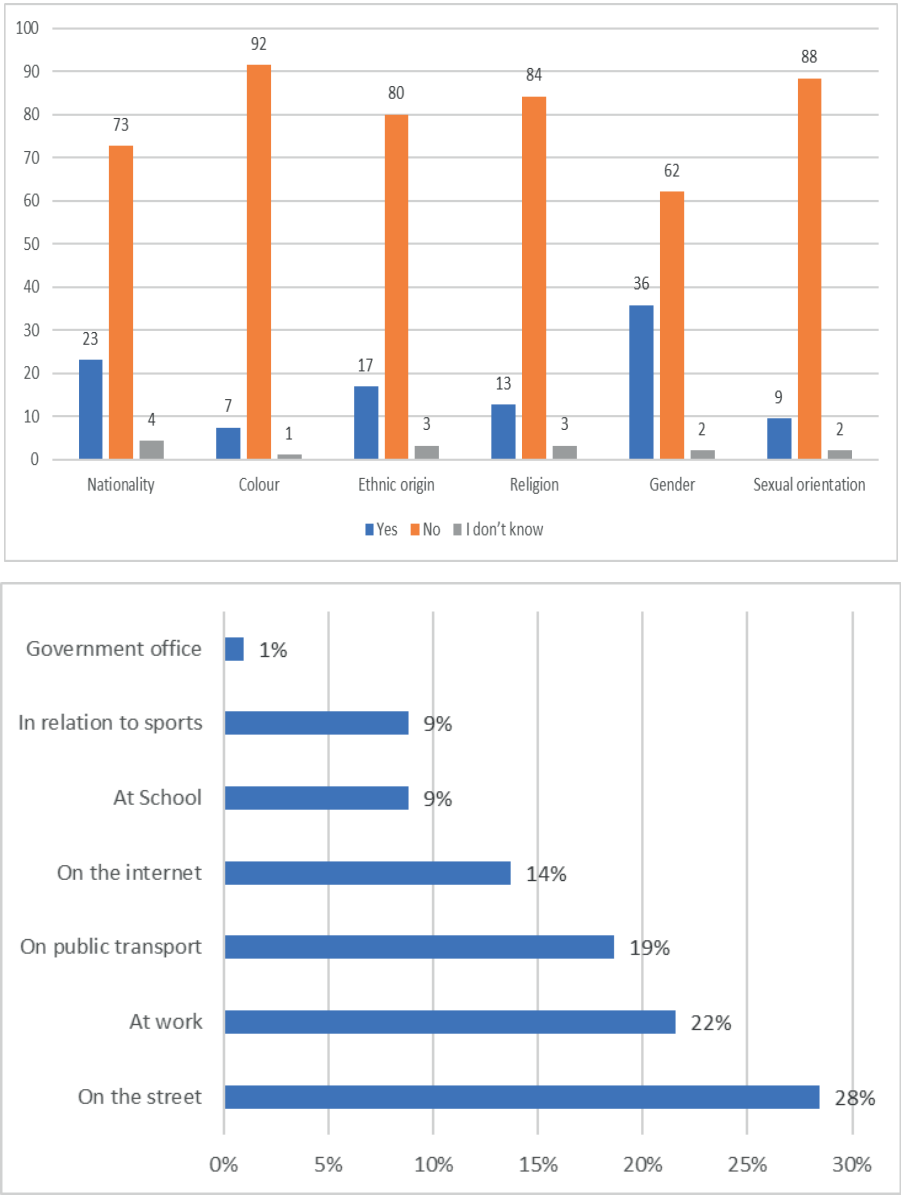


Figure 4: Participants being victim of hate speech

When asked (question 9) why people practice *online* discrimination (fig. 5), participants offered six different, and equally important, reasons: socioeconomic issues (20%), psychological vulnerability (17%), perception of humiliation and discrimination by the local society for ethnic, national, linguistic or religious reasons (17%), normalization of violence (14%), personal causes (divorce, breakup, loss of job) (13%), experience exclusion from their rights (6%), and other reasons (14%). Indeed, it has been recorded elsewhere that socioeconomic issues can explain the surge in online violence (Denti and Faggian 2021), as can the perception of exclusion of rights and discrimination from society (Bouvier 2020).

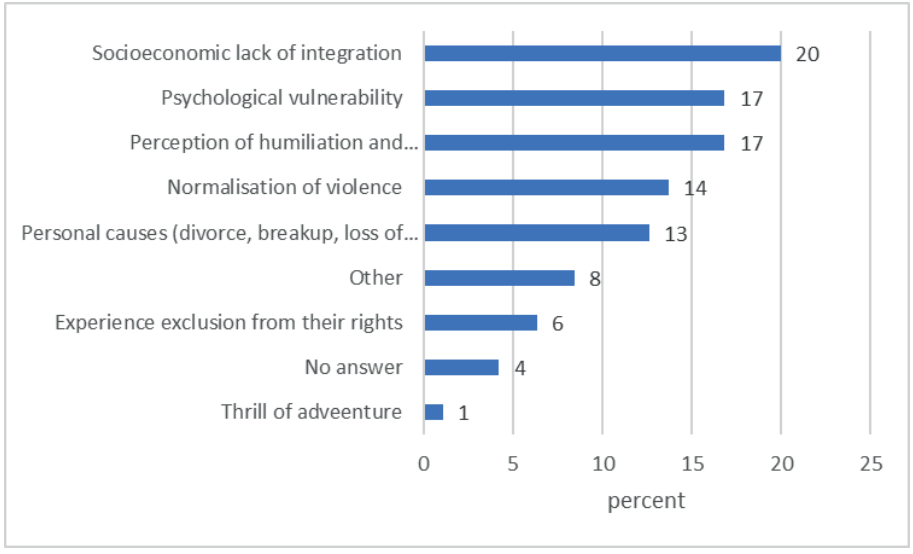


Figure 5: Reasons for practicing online discrimination

When we questioned participants about reporting hate speech, it appears that most people do not intervene (fig. 6) because of lack of trust in the authorities and the belief that actions taken will have no result (53%)[9]. This attitude makes it difficult for authorities to know the extent of such anti-social behavior (Iacob, 2016). Yet when witnessing *online* discriminatory statements, most participants gave one of the following nine answers: *I would feel ashamed, embarrassed or uncomfortable; It would be too much trouble to report it; The incident*

[9] A 2022 European survey focused on corruption found similar figures: "Asked which they thought are the most important reasons people may decide not to report a case of corruption, *the most frequent answer among respondents in Cyprus (53 per cent) was that it would be pointless because those responsible would not be punished.* (our italics)" https://cyprus-mail.com/2022/07/13/94-per-cent-believe-corruption-in-cyprus-is-widespread/

is too common an occurrence to report; Because I'm frightened to be bullied (I would be worried about reprisals from the perpetrator); Because someone else can do it; Because I don't know what to say; I would not know how to report it; Because I don't care, it's part of life, I do not think it is serious enough to report; Because it will have no result, I do not think the Police or authorities would do anything; Because I don't follow online conversations.

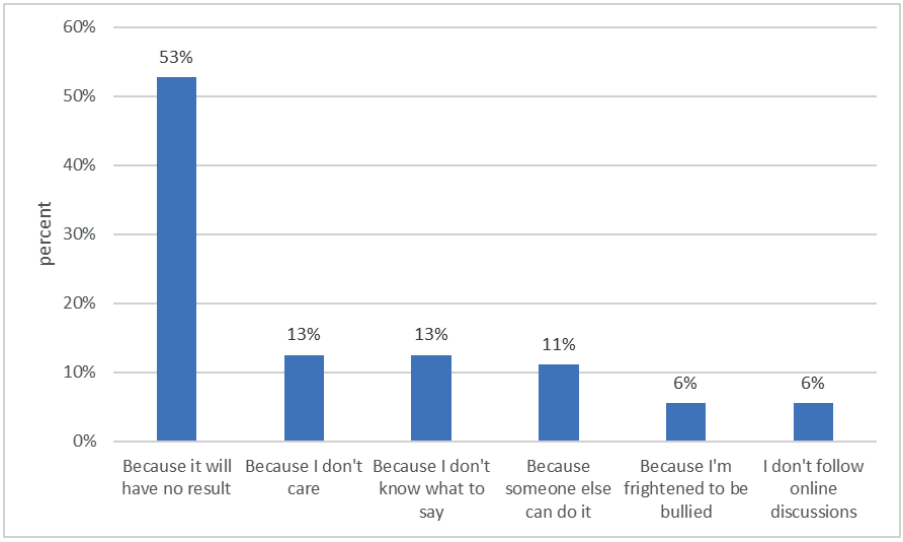


Figure 6: Reasons for not intervening

In those instances where participants reacted to hate speech and intervened (fig.7), the reasons included: to trigger empathy (28%); to offer logical arguments to counter hate speech (23%); to question the ethics of the hateful person (21%); to make a sarcastic comment (17%). These interventions are very similar to the counter speech advocated by Leader Maynard and Benesch (2016), which include persuasion based on emotions (empathy), argumentation (logic, values) and humor (sarcasm). These results will help us choose the memes and tweets that we will use to investigate the effect of counter speech.

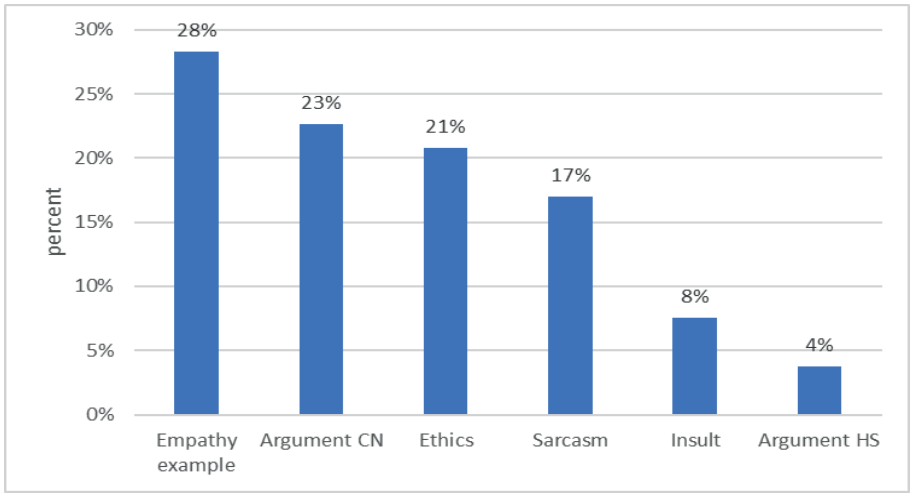


Figure 7: Reactions of participants to hate speech

Additionally, participants witnessed reactions by others against hate speech, mostly by sharing links of videos or images (21%), giving a good explanation as to why this is wrong (16%), pressing the dislike button on Facebook (15%) and giving a logical argument (12%). In the *Other* category, people mentioned they had not witnessed such behavior. Visuals play an important role in online counter speech, which explains our choice of memes. Most counter speech is based on argumentation (explanations and logical arguments) – a result also found in an extensive study of thousands of online comments to hate speech (Baider, in press).

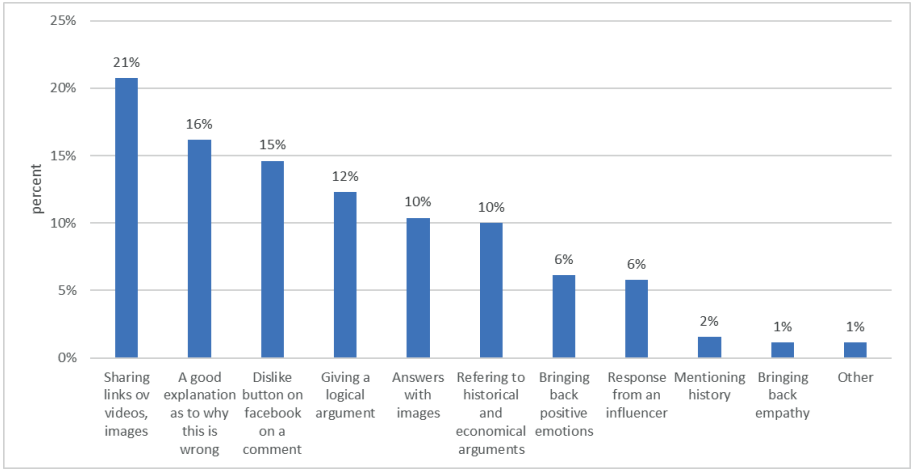


Figure 8: Witnessed reactions to online hate speech

In question 12, participants were asked which response to online discrimination changed their mind; eight arguments against online discrimination were given: *videos, a personal testimony, a logical argument, a good explanation as to why this is wrong , images, bringing empathy, mentioning history, response from an influencer*. Figure 9 shows that the most convincing elements were the presentation of logical arguments (19%), a good explanation why it is wrong to discriminate (19%), and personal testimonies (17%). Videos and images are also used as tools of arguments (11% and 6%, respectively). We note that argumentation seems preferred over persuasion, i.e., using emotions such as trying to bring empathy (15%).

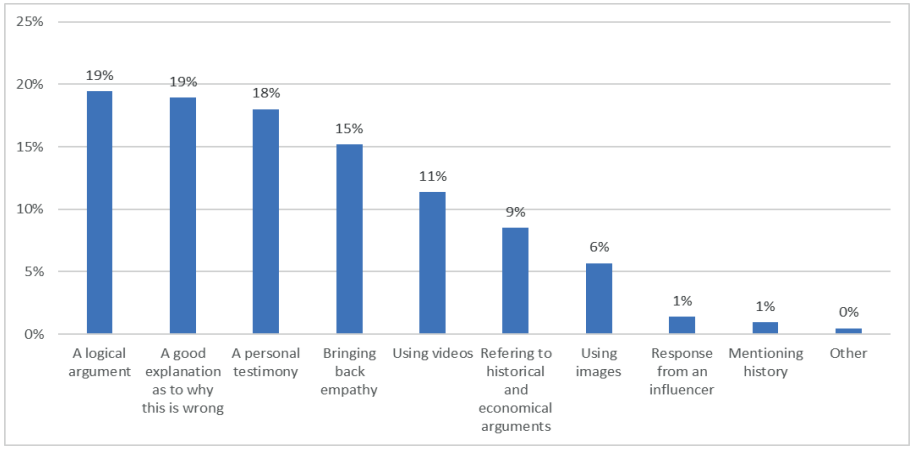


Figure 9: The most convincing arguments against online hate speech

To summarize the results from the questionnaires, we decided to include bio signals to test:

- Tweets based on argumentation (logics and testimonies), and persuasion (emotions such as empathy);
- Images and memes;
- Sexism, since gender-biased discriminatory statements were as prominent, if not more prominent, than racist and homophobic statements.

4. Framing Interviews based on Questionnaire Answers

4.1. Procedure

We undertook 37 interviews with Greek Cypriots in their native language, and three interviews with foreigners in English; all participants had completed the questionnaire prior to the interview. Interviews were guided by responses to the online questionnaires

(presented in section 3.1), with the aim of acquiring a deeper understanding of the answers. Following the same methodology as in our earlier study (Baider 2019), we found that interviews are necessary to contextualize any findings of an online survey.

Participants were interviewed by young Greek Cypriots in 2019, just before the COVID situation preempted any face-to-face interviews[10]. Therefore, all interviews took place face to face and lasted between 15 and 25 minutes, and they were recorded.

For our quantitative analysis, we used corpus linguistics (software SKETCH Engine), while semantics framed the qualitative analysis. Corpus linguistics helps to identify the most frequent co-occurrences in corpora of thousands or millions of words, and uses the Key Word in Context function (KWIC) to understand the textual context of any chosen word.

The concept of frame is here understood as defined by Fillmore:

> any system of concepts related in such a way that to understand any of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text or into a conversation *all of the others are automatically made available.*[11] (1982: 11, our italics)

All interviews given in Greek were compiled and yielded a corpus of 89,385 words. We applied corpus linguistics methodology to investigate these data with the concordance AntConc. In the next section we discuss the frequencies we obtained, and return to the interview extracts to analyze the context for the most important correlations identified in section 3. All interviews were transcribed and translated into English, although we used the Greek transcripts to calculate the frequencies. The words *κοινωνία* 'society' and *μετανάστες* 'migrants' were the most frequent spontaneous words found in the answers.[12]

Κοινωνία 'society' co-occurs with synonyms of the verb *to represent* such as αντικατοπτρίζω, as in αντικατοπτρίζουν την κοινωνία μας 'the figures represent our society'; αντιπροσωπεύω 'represent', as in αντιπροσωπεύει τζαι την κοινωνία 'it represents society'.

Μετανάστες 'migrants' co-occurs with a negative lexical field, e.g., χαμηλότερον 'low', such as in το χαμηλότερον μισθο 'the lowest salary'; υποτιμητικά 'in a derogatory fashion', such as in λαλεί κάποιος

[10] M. Papandreou and G. Constantinou carried out and transcribed the interviews. K. Kyriakides transcribed and translated interviews.

[11] This concept can be also defined, as mentioned by Fillmore himself (1982: 11), with the words *schema, script, scenario* or *cognitive model*.

[12] These frequencies considered the presence of those words in the questions, for example, for the word society, we have 167 occurrences and 53 were in the question asked.

υποτιμητικά προς τους μετανάστες 'someone speaks in a bad way to the migrants'.

4.2. Interview frames

4.2.1. *Κοινωνία* / Society

Some answers complement findings of an earlier study (Baider 2020), in which we pursued interviews with participants who had first completed questionnaires, realizing that interviews were essential to contextualize answers in the questionnaire. In fact, "saving face" (Goffman 1967) may well explain some answers in the questionnaires: although in questionnaires there were certain statements rated unacceptable by a very high percentage, in interviews some felt that this did not reflect reality or their experience.

Especially with regard to (in)tolerance and homosexuality (questions 4-6), participants indicated that the percentages for negative feelings should be even higher, given how conservative Cypriot society is on this issue. For example, see the comment below referring to question 4:

(1) ναι αντικατοπτρίζουν την **κοινωνία** (...) θεωρώ οτι σαν κοινωνία σαν Κύπριοι εν πιο κάτω που 73,7% τζινοι που εν να θεωρούσαν ότι εν είναι αποδεκτό σαν σχόλιο, γιατί είμαστε ακόμα λιο, είμαστε πιο κλιστεί **κοινωνία** σιουρα
'Yes they reflect the society, (...) I consider that as a society like the Cypriot society we should have a number below 73.7% for those who did not consider the comment acceptable, because we are still more, we are a more closed society for sure (...)'

Other participants felt the same way because of their personal experience[13]:

(2) Διαφωνώ, διαφωνώ γιατί τζαι που προσωπικές εμπειρίες
'I disagree, I disagree because of my personal experiences'

Participants suggested that ideology played a role in the results, especially the influence of the Orthodox religion (Karayiannis 2016, Baider 2018):

(3) NS76 Εεεε, πάλε νομίζω ότι εν αντικατοπτρίζουν την κοινωνία μας τούτα τα αποτελέσματα, γιατί εν θεωρώ ότι 73,7% εν το θεωρούν καθόλου αποδεκτό, βασικά εν θέλω να συγκεκριμενοποιήσω, εε, έσσει αρκετό κόσμο που, με κάποιες ιδεολογίες συγκεκριμένες (...)

[13] Some participants belonged to the LGBTQ community and this comment may reflect the experience of LGBT Cypriots.

'Eh, come on, I think that these results do not reflect our society, because I do not think that 73.7% do not consider it acceptable at all, (…). I do not want to specify, but there are enough people who, with because of specific ideologies (…)'

Many respondents also view this conservative attitude to be indicative of the country and the society's backward ways:

(4) C81 Εε, είμαστε, είμαστε ακόμα λίο πίσω που τον κόσμο πας τούντο, part. Εεε, νομίζω, αν τούτη η ερώτηση εγινόταν στο εξωτερικό τζαι όχι στην Κύπρο, ήταν ναν διαφορετικά τα αποτελέσματα. Τη δική μας κοινωνία, ναι αντιπροσωπεύει την.
'Hey, we are, we're still a little behind when you go around the world, part. Eh, I think, if this question had been asked abroad and not in Cyprus, the results would have been different. Our own society, yes it represents it.'

4.2.2. *Μετανάστες* / Migrants

Blatantly racist comments related to *μετανάστες* 'migrants', such as comment 1, were rejected by 93.8% of questionnaire participants; interviewees agreed, with the percentages accepted at face value:

(5) GP77 Συμφωνώ ότι δεν είναι αποδεκτό αυτό το σχόλιο, εμμμ, αν τα αποτελέσματα ήταν διαφορετικά θα ανησυχούσα για την κοινωνία μας...
'I agree that this comment is not acceptable, ummm, if the results were different I would be worried about our society...'

(6) 1978 (…) το θεωρώ έτσι ακραίο, ρατσιστικό, σε πολύ μεγάλο βαθμό.
'I consider it so extreme, racist, to a very large extent.'

The comment most remarked upon in relation to migration was comment 5, which stipulated that "migrants do the jobs Cypriots do not want to do", and represents a negative lexical field co-occurring with the word *μετανάστες*. Some participants negotiated the meaning of the above statement, e.g., *I agree with the percentages but I do not accept the respondent's statement*. Such negotiation is also evident in the percentage range in the questionnaire results (14.5% acceptable, 30% quite acceptable, etc.):

(7) 209 Υπάρχει η έννοια στη Κύπρο, εννοώ, όντος εσχει κάποιες δουλειάς *που οι κύπριοι αρνούνται να καμουν*. (…). Αρά ένταξη πάνω κάτω συμφωνώ με εε τις απαντήσεις αν τζαι προσωπικά θεωρώ ότι εννεν αποδεκτό σαν σχόλιο.
'There is the concept in Cyprus, I mean*, that there are some jobs that Cypriots refuse to do*. (…). Well, OK I more or less agree with the

answers, but I personally think that it is not acceptable as a comment.'

The comment triggered much confusion in terms of whether it was negative or neutral (stating a fact):

(8) 76 Δηλαδή, είτε ποιός το λαλεί, είτε με ποιό τρόπο το λαλεί.
'(it depends) That is, on either who says it, or on the way he/ she says it.'

(9) τούτη η φράση, εε εν αρνητικό. Το να το λέει κάποιος που εε, εν έσσει πρόβλημα με τους μετανάστες.
'This sentence, eh, is not negative. It is said by someone who, eh, has no problem with immigrants.'

In some comments the word *Μετανάστες* is considered in other categories, such as when they are included in the general category of foreigners or more specifically as the Turkish Muslims who invaded their country:

(10) SY43Εεε νομίζω έχει να κάνει με την κουλτούρα μας, την συγκεκριμένη εμάς, διότι θεωρώ ότι σαν κύπριοι είμαστε αρκετά ρατσιστές, σαν σαν άτομα. Εεε και ιδικά με τους μουσουλμάνους τζαι λόγο του προβλημάτος που είχαμεν τελοσπάντων το ότι με το Κυπριακό, και την εισβολή και όλα αυτά
'Well, I think it has to do with our culture, specifically us, because I think that we Cypriots are quite racist, as individuals. Eh, and especially with the Muslims, the reason for the problem we had all along is that with the Cyprus issue, and the invasion and all that'

The macro context, i.e. here, the history of the island, is used to explain some acceptance of racist comments since the country is still in conflict with Turkey and the North is still occupied by Turkish forces (*with the Cyprus issue, and the invasion and all that*). All Muslims and most foreigners will be generalized in a category of people "not welcome": this shift, wherein migrants encompass all foreigners, is revealed in the quotation below:

(11) (...) ίσως, βλέπουμε ξένους, όχι μετανάστες, αλλά γενικά ξένους, σε κάποιες εε δουλειές τέτοιες, που είναι πιο χαμηλών εισοδημάτων
'maybe, we see foreigners, not immigrants, but foreigners in general, in some uh jobs like that, which are lower incomes'

Participants most often explained that foreigners were given these low-paying jobs because of their lack of education (*someone who is not that educated, they will do the jobs that some do not agree to do*):

(12) καποιος μπορει να το δεχτει εεε λόγο της μόρφωσης του και μπορει να

πει οτι κάποιος που δεν ειναι τοσο μορφωμενος εεεεμ, ενα καμν, θα κανει τες δουλειες που δεν καταδέχονται καποιοι να κανου
'someone can accept it because of (lack of) education and we can say that someone who is not that educated, they will do the jobs that some do not agree to do'

However, this is not always the case. Foreign degrees, and even those obtained within the EU, are only accepted with great difficulty; a specific and lengthy process is required, which many deem not worth the time, if they are even aware of what is required. Therefore, they accept work which does not reflect their level of education, e.g., a biologist working in a kiosk as a cashier (personal example).

In summarizing, these few excerpts confirm the results obtained in the questionnaires and contextualise them to justify the results pertaining to migrants. With regard to the homophobic statements, participants felt that the questionnaire results were more positive than in actual fact, and that saving face might explain the many positive comments. Therefore, we could hypothesize that in the psychological experiment homophobic comments will trigger strong reactions; these reactions would also be stronger than when reading racist and anti-migrant statements.

4.2.3. Experience and reactions to offline and online hate speech

This section discusses interviewees' experience of offline and/or online hate speech, although this was most predominant in *online* hate speech:

(13) GP77 Αλλά ναι συμβαίνει, δηλαδή εν εκπλήσσομαι που συμβαίνει, γίνεται, βλέπουμε το ότι γίνεται στην κοινωνία.
'But yes, it is happening, that is, I am not surprised that it is happening, it is happening, we see it happening in society.'

(14) MAP Τζαι ειδικά μέσα στα, Facebook ας πούμε, Instagram, εν κάτι που το παρατηρούμε συχνά νομίζω.
'Well especially in, let's say Facebook, Instagram, something we often notice I think.'

(15) GP77 Ακριβώς, δηλαδή ο κόσμος εστιάζεται στο να κάμνει λεκτική επίθεση, να προσβάλλει, παρά να βάλει κάτω τα επιχειρήματα του, για να πείσει τον άλλο.
'Exactly, that is, people focus on making a verbal attack, insulting, rather than putting down their arguments, in order to convince the other.'

This result does not completely align with a study based on 15,000 manually posted comments on Facebook or YouTube, which

showed that more than 70% of comments use arguments (facts, history, logic, statistics, etc.) (Baider, in press). On the other hand, most of those comments were punctuated by some put-downs such as *dimwits* even after offering solid arguments.[14]

Some mention that the medium, i.e. cyberspace, encourages more insults and verbal attacks and that argumentation has no place. The medium is therefore the problem, a theory proposed since the very first studies focused on verbal aggression and the Internet:

(16) M25 (…), τζαι θεωρώ ότι, βασικά γενικά οι ανθρώποι στο διαδίκτυο (…) οπότε το πρόσωπο που απευθύνεσαι, οπότε εν πιο εύκολο (…) να γράψεις κάτι, παρά να το πεις του άλλου, οπότε θεωρώ απλά γράφουν οι άλλοι τζαι εν αντιλαμβανούμαστε ότι, ούλλοι μας, το τι, το πώς εκλαμβάνεται από το, από το άτομο το οποίο απευθυνόμαστε.
'I think that, basically in general the people on the internet (…) you don't see the other so the person you're addressing, so it's easier (…), to write something, than to say it to the other person, so I think that other people just write and we don't realize that, all of us, how it is perceived by, by the person we are addressing.'

Participants expressed faith in dialogue and believed that using a "gentle" tone in their reactions should yield good results. We found the same in our last study (Baider, in press): to display positive emotions in adversity is the best way "to move forward", as the participant explains below:

(17) 5369 Αν εν κάτι που μπορούμε να βοηθήσουμε την κοινωνία να, αναπτυχθεί να, να πάει μπροστά ας πούμε, εν καλύτερο να αντιδρούμε αλλά ντάξει ήπια, χωρίς προσβολές χωρίς να… καταπατούμε κανενός το δικαίωμα, ή οτιδήποτε
'If in something we can help society to, to develop, to move forward, let's say, it's better to react but gently, without insults without…, trampling on anyone's right, or anything.'

To explain the social reason for people engaging in cyberbullying and verbal violence, interviewees agreed with the results of the questionnaire: the lack of socio-economic inclusion:

(18) GP77 Εμμ, τζαι εγώ **νομίζω** τούτο θα επέλεγα, κοινωνικοοικονομική έλλειψη ένταξης, (…).
'Um, yeah I think this is what I would choose, lack of socio-economic inclusion, (…).'

[14] Such as in the following quotation extracted from the IMsyPP project data base: The real solution to the refugee crisis is – if you don't like refugees, then stop making them refugees by bombing their homelands!!! *Dimwits* (our italics)._IMsyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech) is an EU REC AG cofunded project, ID 875263, http://imsypp.ijs.si/.

For our experiment, therefore, and based on the questionnaire and interview findings, we have:

a. Added sexism to the discriminatory statements to be evaluated;
b. Combined text and image, since in the questionnaire participants were convinced by both.

We thus hypothesize that:

c. Homophobic statements will trigger a greater reaction than racist comments, whether in favor or in disapproval;
d. Text will trigger more reactions that images.

5. The Psychological Experiment

For the bio-signal experiment, we presented both the hate speech statement and the counter speech statement in two different forms: as a meme and as a text as short as a Tweet. We focused on the topics of homophobia, migration, racism and sexism.

5.1. The procedure

The experiment, which tested a small dataset of images and texts for their positive or negative effect on 40 participants[15], was created with the open-source software Open Sesame (Mathôt, Schreij & Theeuwes 2012)[16]. All members of the team chose the stimuli, which were discussed and then agreed.

The entire experiment was developed in an ordered block-design format where blocks of each of the above-mentioned topics were presented in the following order: Homophobia, Racism, Migration, Sexism. Each block opened with a fixation dot in the middle of the screen, which was presented for 995 milliseconds. Next a hate speech narrative was presented until the participant responded, after which a counter narrative was presented. Finally, the participants had to select an answer.

All answers were recorded, as were participants' response time and pulse; at the same time their gaze and movements were recorded on video. The testing procedure was carried out in four stages: in each stage, participants viewed one text and one image related to the questionnaire and interviews (see sections 3 and 4).

- Stage 1 is related to homophobia (text and tweet of hate speech and counter speech; image of hate speech and counter speech), i.e., 6 visuals.

[15] The participants were those who had answered the questionnaire and were interviewed. They received a token fee to encourage their participation in all stages of the study.
[16] https://osdoc.cogsci.nl/

- Stage 2 is related to racism (text *and* tweet of hate speech and counter speech, i.e., 4 stimuli; image of hate speech and counter speech, i.e., 2 stimuli), with a total 6 visuals.
- Stage 3 is related to migration (text *and* tweet of hate speech and counter speech; image of hate speech and counter speech), i.e., 6 visuals.
- Stage 4 is related to sexism (text *and* tweet of hate speech and counter speech; image of hate speech and counter speech), i.e., 6 visuals.

The entire procedure, therefore, comprised 24 stimuli, i.e., 12 for hate speech: 4 categories (homophobia, migration, racism and sexism) with 3 stimuli (meme, tweet, text); the same number for testing counter speech. For 37 participants we have 444 entries for hate speech and 444 for counter speech.

The participants had to rate the stimuli according to a similar scale used for the questionnaire: they were asked to indicate whether they found the content completely acceptable or not at all acceptable in a Likert scale response format ranging from 1 (strongly agree) to 5 (strongly disagree), with 3 being a neutral answer. The Likert scale is a rating scale that quantifies attitudes of the respondents in this experiment on a scale of 1 to 5 (Likert 1932). They selected their answer by tapping on a computer keyboard.

The experiment lasted approximately 15 minutes. For each participant, we sought three types of data: response time, blood pressure and video presentation; as in every quantitative experiment there are a lot of missing values.

5.2. Results and analysis of bio signals

5.2.1. Reaction time

We first collected the reaction time and the rating of the stimuli (as acceptable or not acceptable). Both reaction times and stimulus ratings are catalogued in OpenSesame along with the actual responses on the Likert Scale. At the beginning and at the end of the experiment the researcher catalogued the participants' heart rate with a portable blood pressure monitor.[17]

In Figure 10 we show some indicative data for the participant with ID 1206:

- Column B, 'response', shows the response answers on the 1 to 5 Likert scale;

[17] https://www.omron-healthcare.com/eu/category/blood-pressure-monitors. During the experiment we also recorded a video of the participant in order to analyze facial expression with a face reader.

- Columns C to K give the response time, and column L the total response time;
- Columns N to S give the blood pressure and pulse measurements during the approximately 15 minutes of the experiment.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
ID	response	response_	response_	response_	response_	response_	response_	response_	response_	response_	total_resp	total_resp	High-press	Low-press	Pulse-1	High-press	Low-press	Pulse-2
1206	5	6385	10219	4126	9227	3620	6385	11005	5757	9694	147160	20	118	76	77	119	85	85
1206		11197	10219	4126	11197	3620		6209	5757		110849	16	118	76	77	119	85	85
1206		10942	10219	4126		10942			5662		55497	8	118	76	77	119	85	85
1206		8768	8768	3757							24548	4	118	76	77	119	85	85
1206		8370	8370	3653							12023	2	118	76	77	119	85	85
1206	5	11055	10219	4126	9227	3620	11055	11005	5757	3982	162197	22	118	76	77	119	85	85
1206		9227	10219	4126	9227	3620		11005	5757		131081	18	118	76	77	119	85	85
1206		7179	10219	4126		7179			7826		70502	10	118	76	77	119	85	85
1206	5	7607	10219	4126	9227	3620	7607	11005	5757	5765	175569	24	118	76	77	119	85	85
1206		8150	10219	4126	8150	3620		5414	5757		93443	14	118	76	77	119	85	85
1206		10219	10219	4126							38893	6	118	76	77	119	85	85
1206		3620	10219	4126		3620			5757		79879	12	118	76	77	119	85	85

Figure 10: Indicative data for respondent 1206

Figure 11 shows the stacked bar charts indicating participants' responses in percentages, categorized by the specific stimulus (meme, testimony, tweet) and the hate speech topic (sexism, migration, racism, homophobia). This stacked bar chart offers a clear picture and description of these categorical series. The Likert scale responses vary from 1 (strongly agree) to 5 (strongly disagree), with 3 taking the value of the neutral answer. As can clearly see in the figure, the majority of participants strongly disagreed with the memes, testimonies or tweets of hate speech for all topics examined, with the exception of sexist memes, where we can observe an approximately equal number of participants in agreement and disagreement.

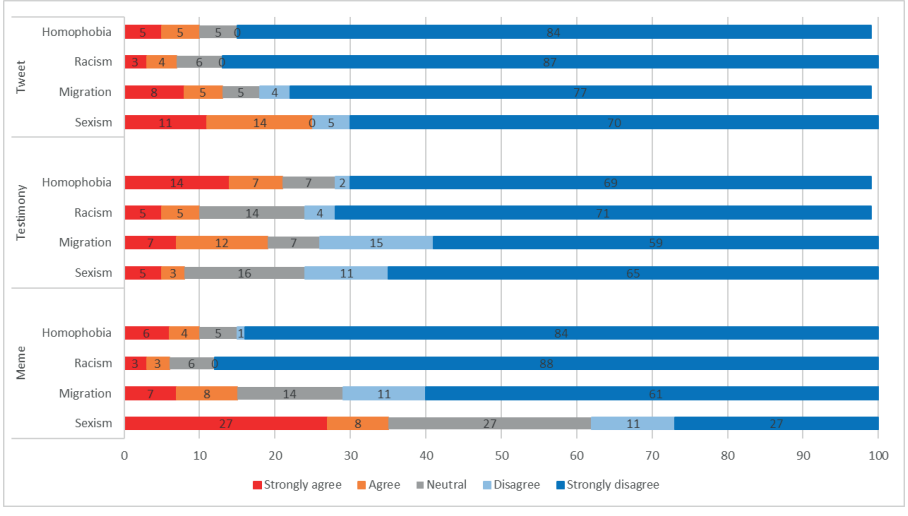


Figure 11: Stacked bar chart for hate speech stimuli and responses

Likert scale series are considered to be ordinal-level data, where each response (1 to 5) is ordered (strongly agree to strongly disagree), but the differences between each number are not meaningful. For this reason, the median and mode statistics are more relevant than the mean.

To test whether there were any significant differences in participant attitudes to the hate speech stimuli, we used non-parametric tests; more precisely, the Kruskal-Wallis test, which is the non-parametric version of the ANOVA test, and which tests for differences in the medians between groups (Kruskal and Wallis 1952). The null hypothesis of the Kruskal-Wallis test is that there are no differences in the medians between groups; in our study, we tested whether the median of the Likert scale responses to hate speech stimuli differed based on the form of communication for each topic. The null hypothesis is rejected if the p-value is less than 0.05.

For the topics of sexism, homophobia and racism, the Kruskal-Wallis H test revealed a statistically significant difference in the median responses to the hate speech stimuli for memes, testimonies and tweets (sexism: N=111, $x^2(2)=16.220$, p=0.000; homophobia: N=444, $x^2(2)=13.536$, p=0.001; racism: N=333, $x^2(2)=12.997$, p=0.002). For the topic of migration, the Kruskal-Wallis H test showed no statistically significant difference in the median responses (agreements) to hate speech memes, testimonies or tweets (migration: N=222, $x^2(2)=4.456$, p=0.108). Subsequent to the Kruskal-Wallis test, we ran the Dunn test (Dunn, 1961) to understand the source of these significant differences in the medians.

These results show that the median keyboard responses for the topic of migration take the value of 5, which corresponds to a strong disagreement. This was true for all three forms of communication: meme, testimony, tweet. On the other hand, for the topics of sexism, homophobia and racism, there were clearly differences in the medians.

The differences in the medians for homophobia and racism stem from testimonies, since the responses are less skewed compared to memes and tweets. The percentage of participants that do not strongly disagree with testimonies of hate speech on homophobia or racism is greater compared to the percentages of participants that do not strongly disagree with memes and tweets of hate speech on homophobia. Nevertheless, the median response value of hate speech stimuli on the topics of homophobia and racism is 5, suggesting strong disagreement.

For sexism, the difference in the medians is traced to memes. The median response for memes is the neutral response of 3, which corresponds to “neither agree nor disagree”, whereas for the other forms of communication (testimony and tweet), the median response is 5, which corresponds to a strong disagreement. This could possibly

suggest that participants have trouble understanding memes on the topic of sexism and they take a neutral approach. This hypothesis can be examined by looking at the differences in the mean response times for each different topic and form of communication. All in all, testimonies and tweets were found to trigger similar answers of non-acceptability of hate speech for all categories (sexism, migration, racism, homophobia). Memes generally triggered answer of non-acceptability for migration, racism, and homophobia, whereas for the case of sexism, they triggered neutral responses[18]. This could indicate a problem in understanding sexist memes, and other memes should test further this idiosyncrasy for sexist memes.

In Figure 12 we see the box plots of the response times and can thus easily visualise the distribution of the series and compare between different categories. Here, we can observe the distribution of response times across the three forms of communication (meme, testimony, tweet) and the four different categories of hate speech (sexism, migration, racism, homophobia). The median response time for each group is indicated by the horizontal line within the bars, and the medians are around 10000 milliseconds for each group. Outliers or extreme values are also visible as dots above the bars, as can be seen for all of the groups.

For the topic of sexism, we can see that there are differences in terms of median values, dispersion and skewness, confirming that sexist memes are evaluated differently. For example, the median value of sexist testimonies is the lowest (approximately 4000 milliseconds), whereas the median value of sexist tweets is the highest (approximately 11000 milliseconds). Furthermore, the distribution of sexist memes is skewed to the right, showing the asymmetry of the distribution of the response times with values higher than the median. This result confirms our suggestion that participants find it difficult to interpret sexist memes. For sexist testimonies and tweets, the distribution appears symmetrical. For the topics of homophobia, migration, and racism, we can observe a number of outliers, although the median value of response times is similar and around 7500, 7000, and 10000 milliseconds, respectively.

[18] A similar picture emerges if we analyze the response means: Sexist memes generated neutral responses, whereas for all other categories and forms of communication the average response was between 4 and 5, indicating disagreements with hate speech.

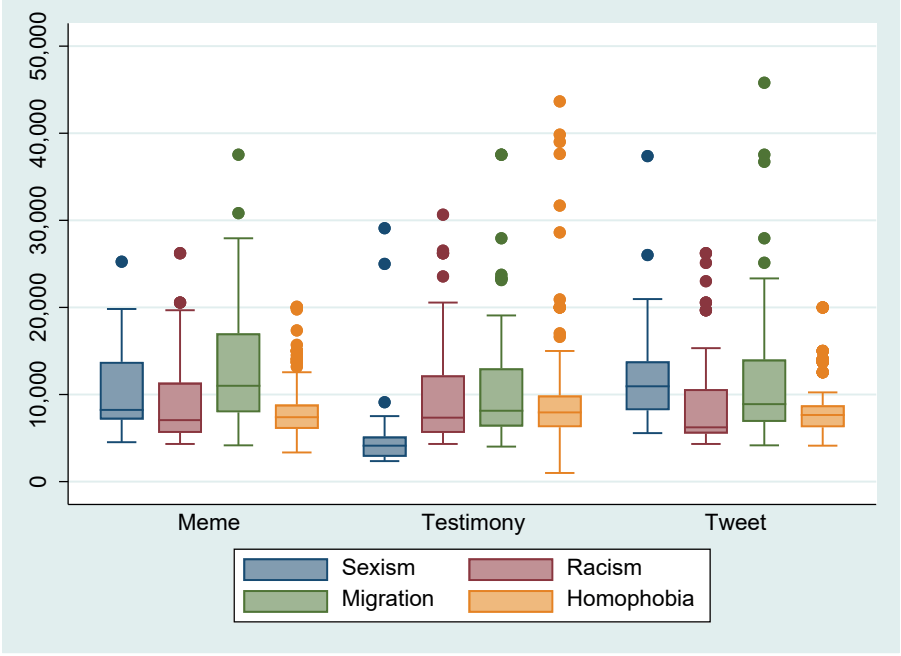


Figure 12: Box plots for hate speech stimuli and response times

We used a parametric measure to test for any significant differences between the response times for the various hate speech stimuli; more precisely, we employed the one-way analysis of variance (ANOVA) test. The null hypothesis of the one-way ANOVA analysis posits that there are no differences in the means across different groups and in our case, we tested whether the mean response times to hate speech stimuli differ based on the form of communication for each topic. The null hypothesis is rejected if the p-value is less than 0.05.

For the topic of sexism, the one-way ANOVA showed that there is a statistically significant difference in the mean response times to the hate speech stimuli for memes, testimonies and tweets (sexism: N=37 for each group, $F(2,108)=15.28$, $p=0.000$). For the topic of homophobia, the one-way ANOVA showed that there is a statistically significant difference in the mean response times to the stimuli of hate speech between memes, testimonies and tweets (homophobia: N=148 for each group, $F(2,441)=4.45$, $p=0.012$). For the topics of migration and racism, the one-way ANOVA showed no statistically significant difference in the mean response times to the hate speech stimuli for memes, testimonies and tweets (migration: N=74 for each group, $F(2,219)=2.13$, $p=0.121$; racism: N=111 for each group, $F(2,330)=0.99$, $p=0.372$). After the one-way ANOVA test, we conducted pairwise

comparisons of means to ascertain the source of these significant differences in the means. These results show that the mean response times for the topics of migration and racism do not differ according to the type of communication (meme, testimony tweet). On the other hand, for the topics of sexism and homophobia, differences in the mean response times were observed. For homophobia and racism, the differences in the mean response time stems from testimonies, as evidenced by the Tukey post hoc test.

Figure 13 displays these mean response times by form of communication (meme, testimony, tweet) and for the different categories of hate speech (sexism, migration, racism, homophobia). As evident in Figure 13, homophobia and racism have similar average reaction time patterns across all three forms of communication, with testimonies taking slightly longer to generate a reaction. These differences are not statistically significant for racism, but are significant for homophobia, based on the one-way ANOVA tests results found above. On the other hand, differences were observed for sexism and migration. Sexist testimonies have the lowest average reaction times, whereas tweets have the highest average reaction times. The results do not support our earlier suggestion that participants might have trouble understanding sexist memes, as the average reaction time of sexist memes is a little over 10000 milliseconds. Lastly, hate speech on migration generates the highest average reaction times of all the categories; nevertheless, these differences are not found to be significant based on the one-way ANOVA test results above.

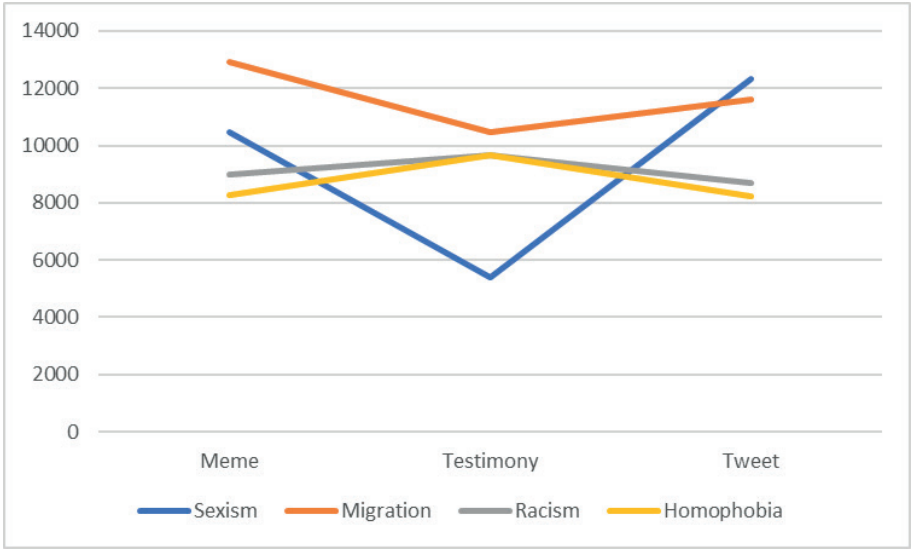


Figure 13: Mean reaction times by stimuli

All in all, for the topics of sexism and homophobia we found differences in the average response times between the different forms of communication, which stem from *testimonies*. Sexist testimonies have a lower average reaction time compared to memes or tweets, whereas homophobic testimonies have a higher average reaction time compared to memes or tweets. We should investigate whether this is true with other memes and tweets. For the topics of racism and migration, there are no significant differences between the different forms of communication; however, hate speech on migration takes a longer time to generate a response, compared to hate speech on racism. Since our sociolinguistic study had shown that racism was readily recognized and rejected, we could hypothesize that the highest time taken to respond may be a sign that the informants do not see the offense in the proposed hate speech segment or that they cannot relate to the counter speech given in the form of testimonies.

5.2.2 Blood Pressure

The blood pressure data (measured before and after the stimuli) was tested to check for any significant difference in the averages. Paired t-tests were used since the blood pressure data corresponded to the same individuals. The results of the paired t-test suggest that the averages of the high blood pressure before (M=112.51, SD=0.55) and after (M=111.7, SD=0.53) the stimuli are not statistically different, i.e., the null hypothesis of zero difference between the two averages of blood pressure is not rejected due to p-value being greater than 0.05 (p=0.11 two-tailed). Additionally, a similar picture emerges when we tested respondents' pulses, both before (M=78.76, SD=0.56) and after (M=79.51, SD=0.57) the stimuli (p-value>0.05 two-tailed). These results suggest that there were no statistically significant changes in the blood pressure or pulse means and these measurements are not relevant as bio-signals for hate speech stimuli.

Based on the above results, we can find no evidence in our experiment to suggest that there are significant changes in bio-signal trackers such as blood pressure or pulse when participants were exposed to hate speech stimuli. Moreover, all participants expressed their (strong) disagreement with the hate speech stimuli and the results suggest that for some topics, testimonies result in different responses (homophobia, racism) and response times (sexism, homophobia). Further investigation is needed to understand the reasons behind the differences between testimonies or other forms of communication. We can also draw some hypotheses about the stimuli we used, and a potential difficulty in determining whether a specific testimony is homophobic or sexist. This difficulty would correlate with the answers

obtained in the questionnaire, which showed that homophobia was much less acknowledged than racism or anti-migrant comments.

6. Conclusions

In this study we used several methodologies to assess reactions to a broad range of hate speech topics, i.e., statements and memes that were racist, homophobic, anti-migrant or sexist. We first conducted a sociolinguistic study: a questionnaire followed by interviews in order to contextualize survey results (Baider 2019); here we noted that participants were likely to be 'saving face' when evaluating sexuality. Next, we ran a psychological experiment, where – as in the sociological study – racist statements were more readily acknowledged as hate speech and rated unacceptable. In fact, the percentage of "strongly disagree" with racist comments is the highest in relation to the other three categories (and for all stimuli – meme/tweet/testimony). Homophobic comments were also readily acknowledged as wrong and rejected (second highest percentage of rejection by adding together disagree + strongly disagree). Statements that were sexist and anti-immigrant were less unanimously evaluated as unacceptable.

These results may be explained by sociocultural factors. Although Cyprus is experiencing a greater sensitivity to racism and discrimination, for example, the Republic of Cyprus has recently organized campaigns against racism in the state schools, the society is overall highly conservative. In light of this, and considering the power and importance of the Orthodox religion, the recorded homophobic attitudes are unsurprising. In our analysis of responses to sexist texts and memes, we must note that the adjective *patriarchal* has been used by several sociologists to describe the Cypriot society (Vassiliadou 2002, Cockburn 2004, Hadjipavlou 2010). We argue therefore that our sociolinguistic and experimental results reflect the socio-cultural context, and we are encouraged to see a correlation in results using the two approaches. We recognize that this study triggers many questions and suggestions, e.g., how to decide which stimuli to choose, whether to test with a questionnaire or an experiment, how to determine a reasonable response time to a content (hate speech), and how to format (difficult comprehension of the meme, for example). Other concerns include: how to best compare the experimental results, which are spontaneous, with interview results that lack this spontaneity; the potential need for a longitudinal study to assess the stability of the results. We can only conclude, with regard to detection of emotions, that if "emotions are complicated, and they develop and change in relation to our cultures and histories" (Crawford 2021), they are also fluid within the same culture, the same community and within the same human being.

References

Almagro, M., Hannikainen, I. and Villanueva, N. (2022), "Whose Words Hurt? Contextual Determinants of Offensive Speech", *Personality and Social Psychology Bulletin*, 48/6, p.937-953.

Assimakopoulos, S., Baider, F. H. and Millar, S. (2017), *Online Hate Speech in the European Union: A Discourse-Analytic Perspective* (1st ed.), Springer, Cham, Switzerland.

Baider F. (2018), "*Go to hell you fagots, may you die*. Framing the LGBTQ subject in online comments", *Lodz Papers in Pragmatics*, 14/1, p. 69-92.

Baider, F. (2019), "Double Speech Act: Negotiating Inter-cultural Beliefs and Intra-cultural Hate Speech among the Youth", *Journal of Pragmatics*, 151, p. 151-166, https://doi.org/10.1016/j.pragma.2019.05.006.

Baider, F. (2020), "Pragmatics lost? Overview, synthesis and proposition in defining online hate speech", *Pragmatics and Society*, 11/2, p. 196-218.

Baider, F. (under evaluation), "Taking down racist metaphors? Impact of Counter speech on (covert) online hate speech", *Politics and Governance*.

Baumgarten, N., Bick, E., Geyer, K., Lund Iversen, D., Kleene, A., Lindø, A. V., Neitsch, J., Niebuhr, O., Nielsen, R. and Petersen, E. N. (2019), "Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS)", *RASK – International journal of language and communication*, 50, p. 87-108.

Bayer, J. and Bard, P. (2020), "Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches", *STUDY*: LIBE Committee, IPOL (Policy Department for Citizen's Rights and Constitutional Affairs), Brussels, European Union, p. 1-169.

Bouvier, G. (2020), "Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice", *Discourse, Context and Media,* 38, p. 1-11.

Brown, A. (2017), "What Is Hate Speech? Part 1: The Myth of Hate", *Law and Philosophy*, 36/4, p. 419-468, doi: 10.1007/s10982-017-9297-1.

Burnap, P. and Williams, M. L. (2016), "Us and them: identifying cyber hate on Twitter across multiple protected characteristics", *EPJ Data Science,* 5, p. 1-15.

Cheng, L., Shu, K., Wu, S., Silva, Y. N., Hall, D. L. and Liu, H. (2020), "Unsupervised cyberbullying detection via time-informed Gaussian mixture model", in *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20). ACM*, online, p. 185-194.

Cockburn, C. (2004), *The Line: Women, Partition and the Gender Order in Cyprus*, Zed Books, London.

Cowan, G. and Hodge, C. (1996), "Judgments of Hate Speech: The Effects of Target Group, Publicness, and Behavioral Responses of the Target", *Journal of Applied Social Psychology*, 26/4, p. 355-374, https://doi.org/10.1111/j.1559-1816.1996.tb01854.x.

Crawford, K. (2021), "Artificial intelligence is misreading human emotion", *The Atlantic*, https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696/

Das, A., Singh Wahi, J. and Li, S. (2020), *Detecting hate speech in multi-modal memes*, arXiv:2012.14891

Davani, A. M., Hoover, J., Atari, M., Kennedy, B., Portillo-Wightman, G., Yeh, L. and Dehghani, M. (2021), "Investigating the Role of Group-Based Morality in Extreme Behavioral Expressions of Prejudice", *Nature Communications*, 12/1, p. 1-13, doi:10.1038/s41467-021-24786-2.

Denti, D. and Faggiana, A. (2021), "Where do angry birds tweet? Income inequality and online hate in Italy", *Cambridge Journal of Regions, Economy and Society*, 14, p. 483-506.

Denzin, N. K. (1978), *The research act: A theoretical introduction to sociological methods*, McGraw-Hill, New York.

Dunn, O. J. (1961), "Multiple comparisons among means", *Journal of the American Statistical Association*, 56/293, p.52-64.

Enarsson, T. and Lindgren, S. (2018), "Free Speech or Hate Speech? A Legal Analysis of The Discourse About Roma On Twitter", *Information and Communications Technology Law,* 28/1, p. 1-18, doi:10.1080/13600834.2018.1494415.

Erjavec, K. and Kovačič, M.P. (2012), "*You Don't Understand, This Is A New War!* Analysis of Hate Speech in News Web Sites' Comments", *Mass Communication and Society*, 15/6, p. 899-920, doi:10.1080/15205436.2011.619679.

Fillmore, Ch. J. (1982), *Frame Semantics. Linguistics in the Morning Calm*, Hanshin Publishing Co, Seoul.

Fischer, A., Halperin, E., Canetti, D. and Jasini, A. (2018), "Why we hate", *Emotion Review*, 10/4, p. 309-320.

Fortuna, P. and Nunes, S. (2018), "A Survey on Automatic Detection of Hate Speech in Text", *ACM Computer Surveys*, 51/4, p. 1-30, https://doi.org/10.1145/3232676.

Gagliardone, I., Gal, D., Alves, T. and Martinez, G. (2015), *Countering Online Hate Speech*, United Nations Educational, Scientific and Cultural Organization, Paris.

Goffman, E. (1967), "On Face-Work. An Analysis of Ritual Elements in Social Interaction", in *Ders: Interaction Ritual*, Doubleday, New York, p. 5-45.

Guillén-Nieto, V. (2020), "Defamation as a Language Crime. A Sociopragmatic Approach to Defamation Cases in the High Courts of Justice of Spain", *International Journal of Language and Law*, 9, p. 1-22.

Guillén-Nieto, V. (2022), "What else can you do to pass?", in J. Giltrow *et al.* (eds), *Legal meanings. The Making and Use of Meaning in Legal Reasoning*, De Gruyter, Berlin, p. 31-55.

Hadjigeorgiou, N. (2020), *Report on the status of foreign domestic workers in Cyprus* https://equineteurope.org/wp-content/uploads/2021/01/Cyprus-Domestic-Workers.pdf

Hadjipavlou, M. (2010), *Women and Change in Cyprus: Feminism, Gender and Conflict*, I.B. Tauris, London/New York.

Herring, S. (2002), "Computer-mediated communication on the Internet", *Annual Review of Information Science and Technology*, 36, p. 109-168.

Iacob, O. (2016), "Hate Crime and Hate Speech in Europe: Comprehensive Analysis of International Law Principles, EU-Wide Study and National Assessments", *Preventing Redressing and Inhibiting Hate Speech in New Media*, Fundamental Rights and Citizenship Programme of the European Union, Bucharest.

Kamenou, N. (2011), "Queer in Cyprus: national identity and the construction of gender and sexuality", in L. Downing and R. Gillett (eds), *Queer in Europe* (Queer Interventions Series), Ashgate, Surrey and Burlington, p. 25-40.

Karayiannis, S. S. (2016), "Through the narrow straits: Researching homophobia and sexual oppression in Cyprus", in C. N. Phellas (ed.), *Researching Non-Heterosexual Sexualities*, Routledge, London/New York, p. 187-200.

Karlekar, S. and Bansal, M. (2018), "SafeCity: Understanding diverse forms of sexual harassment personal stories, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18)*, ACL, Brussels, p. 2805-2811.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P. and Testuggine, D. (2020), "The Hateful Memes Challenge: Detecting hate speech in multimodal memes", in *Proceedings of the Annual Conference on Neural Information Processing Systems* (NeuIPS '20), online.

Kruskal, W. H. and Wallis, W. A. (1952), "Use of ranks in one-criterion variance analysis", *Journal of the American Statistical Association*, 47/260, p. 583-621.

Leader Maynard, J. and Benesch, S. (2016), "Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention", *Genocide Studies and Prevention: An International Journal (GSP)*, 9/3, p. 70-95.

Likert, R. (1932), "A technique for the measurement of attitudes", *Archives of Psychology,* 140, p. 5-55.

Masud, S., Bedi, M., Khan, M.A., Akhtar, M.S. and Chakraborty, T. (2022), "Proactively Reducing the Hate Intensity of Online Posts via Hate Speech Normalization", arXiv preprint arXiv:2206.04007.

Mathôt, S., Schreij, D. and Theeuwes, J. (2012), "OpenSesame: An open-source, graphical experiment builder for the social sciences", *Behavior Research Methods*, 44/2, p. 314-324, doi: 10.3758/s13428-011-0168-7.

Neitsch, J. and Niebuhr, O. (2020), "On the role of prosody in the production and evaluation of German hate speech", in *Proceedings of the 10th International Conference on Speech Prosody*, Japan, Tokyo, p. 710-714.

Paz, M.-A., Montero-Díaz, J. and Moreno-Delgado, A. (2020), "Hate Speech: A Systematized Review", *SAGE Open*, 10/4, p. 1-12, doi: 10.1177/2158244020973022.

Poria, S., Cambria, E., Hazarika, D., and Vij, P. (2017), A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks", in *Proceedings of the 26th International Conference on Computational Linguistics - COLING 2016*, p. 1601-1612.

Rasaq, A., Udende, P., Abubakar, I.Y. and La'aro, O. A. (2017), "Media, Politics, and Hate Speech: A Critical Discourse Analysis, *E-Academia Journal*, 6/1, p. 240-252.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A. and Choi, Y. (2020), "Social bias frames: Reasoning about social and power implications of language", in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*, ACL, p. 5477-5490, online.

Srba, I., Lenzini, G., Pikuliak, M. and Pecar, S. (2021), "Addressing Hate Speech with Data Science: An Overview from Computer Science Perspective", in S. Wachs *et al.* (eds), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen*, Springer VS, Wiesbaden, p. 317-336, https://doi.org/10.1007/978-3-658-31793-5_14.

Stephan, W. G., Ybarra, O. and Bachman, G. (1999), "Prejudice Toward Immigrants", *Journal of Applied Social Psychology*, 29/11, p. 2221-2237, https://doi.org/10.1111/j.1559-1816.1999.tb00107.x.

Švec, A., Pikuliak, M., Šimko, M. and Bieliková, M. (2018), "Improving moderation of online discussions via interpretable neural models", in *Proceedings of the second workshop on abusive language online-ALW2*, Association for Computational Linguistics, Brussels, p. 60-65.

Vassiliadou, M. (2002), "Questioning nationalism: the patriarchal and national struggles of Cypriot women within a European context", *European Journal of Women's Studies*, 9/4, p. 459-482.

Wachs, S., Koch-Priewe, B. and Zick, A. (eds) (2021), *Hate Speech – Multidisziplinäre Analysen und Handlungsoptionen. Theoretische und empirische Annäherungen an ein interdisziplinäres Phänomen*, Springer Fachmedien, Wiesbaden.

Waseem, Z. and Hovy, D. (2016), "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter", in *Proceedings of the NAACL Student Research Workshop*, ACL, San Diego, p. 88-93.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R. (2019), "Predicting the type and target of offensive posts in social media", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '19)*, ACL, Minneapolis, p. 1415-1420.

Zhang, Z. and Luo, L. (2019), "Hate speech detection: A solved problem? The challenging case of long tail on Twitter", *Semantic Web*, 10/5, p. 925-945.