Queer NLP: A Critical Survey on Literature Gaps, Biases and Trends

Sabine Weber[1,2,*] **Angelina Wang**[1,3,†] **Ankush Gupta**[1,4,5,†] **Arjun Subramonian**[1,†]
Dennis Ulmer[1,6,†] **Eshaan Tanwar**[1,7,†] **Geetanjali Aich**[1,8,†] **Hannah Devinney**[1,9,†]
Jacob Hobbs[1,17,†] **Jennifer Mickel**[1,10,†] **Joshua Tint**[1,11,†] **Mae Sosto**[1,12,†]
Ray Groshan[1,13,†] **Simone Astarita**[1,†] **Vagrant Gautam**[1,14,†] **Verena Blaschke**[1,15,16,†]
William Agnew[1,17,†] **Wilson Y Lee**[1,18,†] **Yanan Long**[1,19,†]

[1]Queer in AI [2]University of Bamberg [3]Cornell Tech [4]IIIT-DELHI [5]PayGlocal
[6]ILLC, University of Amsterdam [7]University of Copenhagen [8]University of Massachusetts Amherst
[9]Linköping University [10]EleutherAI [11]Arizona State University
[12]Centrum Wiskunde & Informatica (CWI) [13]University of Maryland, Baltimore County
[14]Heidelberg Institute for Theoretical Studies (HITS) [15]LMU Munich [16]MCML
[17]Carnegie Mellon University [18]HubSpot [19]StickFlux Labs

Abstract

Natural language processing (NLP) technologies are rapidly reshaping how language is created, processed, and interpreted by humans. With current and potential applications in hiring, law, healthcare, and other areas that impact people's lives, understanding and mitigating harms towards marginalized groups is critical. In this survey, we examine NLP research papers that explicitly address the relationship between LGBTQIA+ communities and NLP technologies. We systematically review all such papers published in the ACL Anthology up until February 2026 (n=122), to answer the following research questions: (1) What are current research trends? (2) What gaps exist in terms of topics and methods? (3) What areas are open for future work? We find that while the number of papers on queer NLP has grown within the last few years, most papers take a reactive rather than a proactive approach, focusing on shortcomings of existing systems rather than creating new solutions. Our survey uncovers many opportunities for future work, especially regarding stakeholder involvement, intersectionality, interdisciplinarity, and languages other than English. We also offer an outlook from a queer studies perspective, highlighting understudied topics and blind spots in the harms addressed in NLP papers. Beyond being a roadmap of what has been done, this survey is a call to action for work towards more just and inclusive NLP technologies.

webersab/queer_NLP_bib

*First author, please send correspondence to sabine.weber@uni-bamberg.de.

†Equal contributions, alphabetical order by first name

1 Introduction

In this survey, we examine the state of NLP research as it relates to LGBTQIA+ topics, which we call Queer NLP (for an in-depth definition, see Section 3.1). We use the term *queer* as a shorthand for the broad spectrum of identities represented under the LGBTQIA+ umbrella (see Glossary, §A), while recognizing the diverse and heterogeneous experiences within this community. Although there has been a significant amount of work in NLP on bias more broadly (Stanczak and Augenstein, 2021; Field et al., 2021; Goldfarb-Tarrant et al., 2023), we foreground the queer community by surveying papers that discuss the distinct social impacts queer people face in the context of NLP technologies.

As evidenced by an increasing number of articles and its first archival event for ACL (Queer in AI @ NAACL 2025; Pranav et al., 2025), Queer NLP is growing as a topic of study, which necessitates synthesizing and systematizing the knowledge in this subfield to understand current trends, identify gaps, and assess promising future directions. These objectives are critical for NLP to meaningfully engage with queer communities and their linguistic realities to foster more inclusive and socially responsible systems. Therefore, we systematically survey all Queer NLP papers from the ACL Anthology. We annotate these papers for tasks, types of data, queer groups and harms they focus on, as well as their motivations, methods, and engagement with intersectionality and stakeholders.

Our main findings (cf. §4) are:

1. **Trends:** While there is increased interest in Queer NLP, approaches are often guided by the affordances of current models, focusing on pronouns and hate speech as prominent

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2625–2649, 2026. https://doi.org/10.1162/TACL.a.818
Action Editor: Kai-Wei Chang. Submission batch: 10/2025; Revision batch: 5/2026; Published 10/2026.
©2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

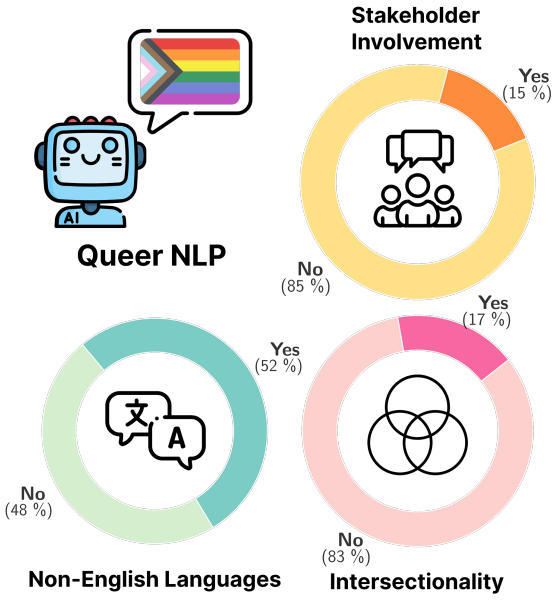


Figure 1: We find that the majority of Queer NLP papers published in the ACL Anthology disregard intersectionality and omit stakeholders. Slightly more than half of all papers include languages other than English. These are descriptive findings, and we provide recommendations in §5.

topics (§4.1) and template-based approaches and data augmentation as prominent methods (§4.2).

2. **Gaps:** Approaches rarely involve stakeholders or take an intersectional view. Nearly half of all papers focus only on English (Figure 1; §5).

3. **Future directions:** Beyond closing these methodological gaps, we call on the field to address larger structural gaps, such as dynamic evaluations which allow for temporal and contextual differences, ongoing stewardship of language technologies by the affected communities and a stronger involvement with other disciplines like sociology, gender studies and queer studies (§5).

Natural language processing (NLP) is a field concerned with enabling computers to analyze, understand, and generate human language. As NLP technologies become increasingly integrated into everyday life, users' direct interaction with them has expanded. It is critical that these systems function well, do not reinforce social harms, and maintain the privacy of all users, including queer individuals, who have historically been excluded from technological development and research (Scott et al., 2017; Organizers of Queer in AI et al., 2023).

2 Related Work

The concepts of gender and sexual orientation are central to Queer NLP and have been examined in a number of prior studies. Devinney et al. (2022) survey 200 articles on gender bias in NLP to examine how the field conceptualizes gender, Cao and Daumé III (2020) review 150 articles on gender in co-reference resolution. Hobbs (2025) surveys 55 papers on sexuality bias in NLP, while Zhou (2024) offers a broad overview of queer harms and structural barriers in NLP. Our work extends these previous approaches with an in-depth survey of the field, spanning all of the named topics and adding ACL anthology papers from different tasks as well, e.g. from machine translation and speech tasks.

Outside of Queer NLP, several papers survey research on social groups and issues in NLP, including gender bias (Stanczak and Augenstein, 2021) and racial bias (Field et al., 2021). Others survey specific terms, aiming to investigate how they are used in the field, e.g., "low-resource" (Nigatu et al., 2024), "intersectionality" (Ovalle et al., 2023b; Wang et al., 2022), and "democratization" (Subramonian et al., 2024). To our knowledge, our paper is the first to systematically survey Queer NLP papers within the ACL Anthology.

3 Methodology

3.1 Definition of Queer NLP

For our purposes, "Queer NLP" describes published research work that examines the relationship between the LGBTQIA+ community and NLP technologies. Queer NLP examines how language technologies operationalize concepts related to gender and sexual orientation.

Queer NLP is not simply the application of general bias detection and mitigation techniques to a demographic group. While Queer NLP draws on and contributes to broader fairness and bias methodologies, it addresses modeling challenges that do not easily translate to other groups (Wang, 2025). It can draw on findings from disciplines like gender studies and queer studies and span multiple levels of linguistic representation. Queer NLP includes phonetic phenomena such as the transcription of queer speakers' voices, syntactic phenomena such as pronoun usage, pragmatic phenomena such as hate speech and reclaimed slurs, and discourse-level phenomena such as code-switching and the

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

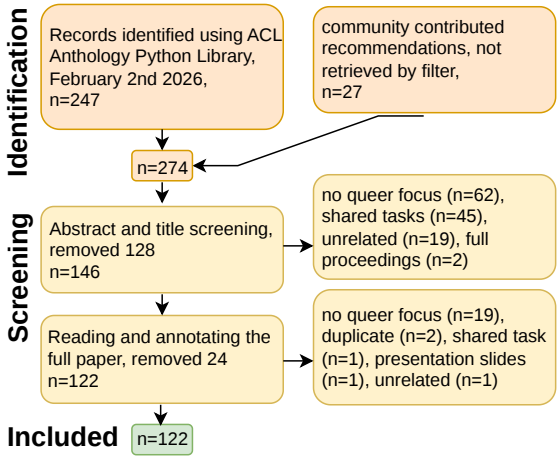


Figure 2: Selection process and filtering for the used papers

co-construction of identity through interaction over time. Consequently, Queer NLP constitutes a distinct and multifaceted area of inquiry focused on phenomena that are linguistically diverse, context-dependent, and resistant to simplification.

3.2 Paper Selection

Selection Criteria. We intentionally limit our scope to papers that place queer people at the center of their inquiry. In this spirit, we only include NLP research that explicitly engages with queerness as a core axis. We exclude work where the engagement with queerness is not a substantial part of the paper, e.g., when queer identity terms are used only as an illustration for a wider phenomenon, queer users are only one of many examined demographic subcategories, or where gender inclusivity is only applied to a binary conceptualization of gender.

We also constrain our paper selection to papers published in the ACL Anthology. While NLP papers are published in other venues as well, the ACL Anthology represents a significant proportion of the discourse in NLP. We do explicitly include papers published at workshops, because research on marginalized communities is, despite its valuable contributions, often not represented at mainstream venues.

Selection Method. To operationalize these principles we take the following steps (Figure 2): We use the ACL Anthology Python library[1] to retrieve papers that feature one or more keywords from a predefined list of queer-specific terms (collected by us) in either their title or abstract. We hypothesize that papers that focus on queer topics would name them prominently. The list contains a subset of the terms used by Taylor et al. (2024): *queer, aromantic, gender non*, lgbt*, agender, glbt, lesbian, gay, bisexual, transgender*, but also words related to systemic oppression such as *heteronormative*, *homophobia*, and *transphobia* and words describing queer-specific harm mitigation such as *gender inclusive* and *gender-fair*. The full list can be seen in Appendix E. The retrieval was performed on February 2nd 2026, covering all papers in the ACL Anthology previous to this date. This yields 247 papers.

We also allow community contributions provided by us, the authors of this survey, covering papers that could fit our selection criteria but were not retrieved by our search term list. Examples of this category are a paper concerned with neopronouns (Gautam et al., 2024a), which are often used by non-binary people, or a paper on a chatbot for education about HIV/AIDS (Brixey et al., 2017), which disproportionally affects queer people. This step contributes 27 papers.

We then apply two layers of filtering to check if the papers match the selection criteria. The first filter consists of the manual reading of the titles and abstracts of the papers. This step removes 128 papers, most of them because they lack a queer focus or are unrelated (e.g., when the word 'non-binary' describes the property of a graph and not a gender identity). We also remove papers that are submissions to shared tasks, as these would inflate paper numbers while covering the same data set and task. The second filtering is performed during the full-paper annotation step (§3.3) and removes 24 papers. Our final survey covers 122 papers.

Questions about whether a paper counts as "Queer NLP" are resolved through discussion among us, the authors, using the outlined selection criteria. While we make statistical claims only about the papers contained within the ACL Anthology, our combined expertise brought up many papers that are preprints or published in other venues, which we cite when their perspective is necessary to gain a full understanding of the field. We present all papers as a living resource in an open-source repository, which also contains a file with our annotations.[2]

[1] acl-anthology.readthedocs.io/py-v1.1.0

[2] https://github.com/webersab/queer_NLP_bib

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

3.3 Annotation Process

We annotate each paper in our survey along several axes. First, we annotate the NLP task (e.g., question answering), the motivation, and the method. Second, we annotate which queer groups are explicitly named and what kinds of harm are addressed (if mentioned) and limitations named in the paper.

We focus on three themes: language diversity, intersectionality and stakeholder involvement. We document the language (see Figure 3) as well as the geographical region and genre of the data used.

We also annotate whether intersectionality is explicitly addressed. Intersectionality is a critical framework combining inquiry and praxis to understand how intersecting systems of power reinforce social inequalities (Collins and Bilge, 2020; Crenshaw, 1989). In practice, this could include overlapping forms of discrimination, such as harms uniquely affecting Black trans women that differ from those faced by white trans women or Black cis women. We examine how papers account for this complexity and contextualize queer bias and harms within broader systems of power.

Our third focus theme is the explicit involvement of any stakeholders. While work in NLP might not always require stakeholder involvement, there is a growing awareness within the field for the necessity of this approach when dealing with marginalized groups (see Caselli et al., 2021 and Dinkar et al., 2026).

Each paper is read and annotated once, but to calculate inter-rater reliability we randomly select 11 papers (9% of all collected papers) to be annotated by a second person. We report high reliability across our three focus themes of language diversity, intersectionality and stakeholder involvement, with perfect agreement for language diversity and 90.09% agreement for the other themes (Cohen's Kappa = .62 for intersectionality and .79 for stakeholder involvement) .

4 Findings: Common Themes

Surveying Queer NLP research, we find common trends in what problems are targeted. We also find trends in the methods used. In this section we first look at phenomena that concern queer communities that are reflected in language (4.1) and then examine the methods used both for analysis and mitigation (4.2). While we focus on papers retrieved in our survey, sometimes papers from outside the ACL Anthology can provide useful insights. To distinguish these papers we add a * to their citation in this section.

4.1 Objects of Study

4.1.1 Pronouns

Pronouns are closely connected to the queer community because they function as a linguistic mechanism through which individuals signal, negotiate, and affirm gender identity in languages where pronouns carry gender information. Gender-neutral pronouns (especially in languages where gendered pronouns are considered the default) and neopronouns feature prominently in our survey. Underrepresentation of specific pronouns (e.g. English "xier") and pronoun–noun combinations (e.g. "her wife") in training data has been shown to impair performance in coreference resolution and translation tasks (Stewart and Mihalcea, 2024; Friðriksdóttir, 2024). 14 of our papers (11%) investigate these issues.

Papers evaluating coreference resolution highlight the harms of misrecognizing people's genders, disconnecting someone's name from their pronouns, or linking their name to a different pronominal chain. Most papers name trans and non-binary people as the targets of potential harm (Dev et al., 2021), while Cao and Daumé III (2020) state that cis people can also be harmed by misgendering, and Gautam et al. (2024b) argue from a standpoint of universal design (i.e., designing systems to be usable by the widest range of people) without naming specific communities.

The machine translation (MT) community looks at this problem in regard to translating between languages that mismatch in the gender marking of terms. This work is often found under the term "gender-fair language." and comprises both gender-inclusive and gender-neutral language (Paolucci et al., 2023). At first glance, it is often unclear which papers rely on a binary conception of gender and which explicitly incorporate gender-neutral terms used by trans and non-binary people. Several papers focus on neopronouns: For example, Piergentili et al. (2024) introduce NEO-GATE, an extension of the GATE benchmark (Rarrick et al., 2023*), for translation from English to Italian, and Piergentili et al. (2025) adapt chain-of-thought prompting to improve gender-neutral translation. Lauscher et al. (2023) incorporate nounself, emojiself, numberself, and other neopronouns in addition to gender-neutral and gendered pronouns while

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

evaluating MT models.

Another way to create gender-fair texts is rewriting and post-editing: Lardelli and Gromann (2023) recruit translators to post-edit English to German machine translations into gender-fair language, while Veloso et al. (2023) create a rule-based and a neural-based rewriter for Portuguese. Amrhein et al. (2023) use round-trip translation to create pairings of unbiased texts with biased translations to train a de-biasing rewriting system.

Recommendations Resolution of neopronouns or non-normative combinations of pronouns and nouns is one exemplary case in which human performance easily surpasses state-of-the-art machine learning models, serving as a litmus test for when a machine learning model is memorizing surface forms rather than learning underlying rules. This makes it a phenomenon that is easily measurable and therefore shows up often in NLP publications. Nevertheless, how downstream performance affects queer users remains under-examined. Other examples of research gaps in this area are the treatment of context-based switching between pronoun sets and the usage of several different pronoun sets for the same person. A way forward is increased stakeholder involvement, as described by Gromann et al. (2023) who study how trans and non-binary people are impacted by machine translation and what their wishes for better systems are.

4.1.2 Hate Speech and Slurs

In recent years, detecting transphobic and homophobic hate speech has become a prominent focus, with numerous shared tasks addressing this challenge across multiple languages (Chakravarthi et al., 2022, 2023, 2024; Nozza et al., 2023; Bel-Enguix et al., 2023*; Gómez-Adorno et al., 2024*; Jiménez-Zafra et al., 2023*; García-Baena et al., 2024*). 34 of the papers in our survey (28%) examine hate speech.

When trained on data stripped of its situational context, systems frequently misclassify identity terms as slurs or negative expressions. This leads to problems in hate speech classification, where systems over-classify instances with queer terms as offensive, while not flagging offensive text that lack these words. Several papers show that models regularly misclassify words like *gay, trans,* or *queer* as toxic, even when used neutrally or positively (Zueva et al., 2020; Sahoo et al., 2022). Some papers show that text classification systems are also biased against language used among queer people (Wu and Hsieh, 2017; Ramesh et al., 2022) and references to queer identities (Tint, 2025; Ungless et al., 2023; Zhang et al., 2020; Zueva et al., 2020).

Despite the prevalence of hate speech in a diversity of languages, a relatively small number of papers study languages other than English. Work has been done on Mexican Spanish (Vásquez et al., 2023), Russian (Zueva et al., 2020), Malayalam and Tamil (Sharma et al., 2023*), and Gujarati, Kannada, and Telugu (Kumaresan et al., 2024). Locatelli et al. (2023) study homotransphobic speech on Twitter in seven languages, concluding that current automated moderation systems are insufficient and have large differences across languages.

Recommendations Online content moderation is one of the spaces where queer users can directly feel the impact of language technologies by being silenced, censored or targeted, which makes it an impactful field of study. In the case of hate speech, who is speaking to whom is just as important as what is being said, but there is a gap in research on in-group and out-group language (one notable exception being Draetta et al. (2024)). However, detecting in-group language can be a double-edged sword, as machine learning systems that recognize in-group usage of queer language to prevent censorship can be used to expose queer people who use coded language to escape detection. We recommend the community focus on the tension between detecting in-group language use and safeguarding those who rely on it for privacy and safety. Also, current methods have focused primarily on flagging harmful speech and less on promoting counterspeech. Notable exceptions are works on hope speech which have started to appear in the NLP ethics community, including in Queer NLP (Pofcher et al., 2025).

4.1.3 Stereotypes

Besides the category of hate speech, queer people are represented in the public discourse in many different ways. Stereotypes of queer people are propagated across social and traditional media, their presence or absence reflecting localized climates of acceptance or hostility. When stigmatizing stereotypes become training data, they can be reproduced by NLP systems, resulting in harmful biases that surface in downstream applications. In this sense, stereotypes can be studied both in the underlying training data and in the analysis of LLM outputs,

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

which opens its own field of study that is distinct from the examination of hate speech and slurs.

Some works use information extraction tools to evaluate texts written about the queer community. Locatelli et al. (2023) and Andersen et al. (2024) analyze English and Mexican Spanish homotransphobic speech on Twitter, respectively, noting a decrease in the use of many derogatory words over time, while the overall vocabulary for queer terms has increased. CH-Wang and Jurgens (2021) also study Twitter and Reddit data, finding that the use of gender-neutral expressions (e.g., partner, spouse, folks) has risen steadily, with the sharpest uptake among heterosexual, liberal, and higher-socioeconomic-status users. Hicks et al. (2016) use Twitter data to find terms for gender expression that do not appear in the US National Transgender Discrimination Survey to improve survey design. These papers demonstrate the potential of NLP tools in analyzing the use of queer terms in society more broadly, leading to increased knowledge of the popular discourse about queer people.

When it comes to queer representation in the outputs of LLMs, there has been a wide application of template-based methods to elicit stereotypes, which we cover in more detail in §4.2.

An important aspect of stereotypes is their changeability. Zhou (2024) draws attention to the fact that understandings of queerness vary both across cultures and over time. Given the fluid nature of "queerness" as used by queer people and noted by philosophers (Butler, 2020)*, models of anti-queer bias must be equally fluid to account for different dataset contexts. For instance, inoffensive or clinical terms from the past, such as "transvestite", can gain negative associations through pejoration (Finkbeiner et al., 2016)*.

Representation of queerness in NLP technology would hence require constant rewriting by the affected community, similarly to the slow and constant change of language in different communities of practice, which the affordances of current NLP technologies make hard to imagine.

Recommendation It remains important to examine how queer people are discussed online and how ongoing language change shapes the terms associated with them. LLM outputs can be viewed as a reflection of such discourse, while also revealing these systems' tendency to internalize latent social patterns, including hetero- and cisnormativity. Zhou et al. (2025)* point out that LLMs are only capable of reproducing discourse about social groups, rather than capturing how social identities are enacted through language. Looking into this topic is in line with the study of in-group language and would shine a light not only the most overt manifestations of bias but also on subtle and context dependent patterns of normativity and erasure.

4.1.4 Underrepresentation

In many curated datasets, queer data are underrepresented or even filtered in fear of being offensive (Dodge et al., 2021). Lack of queer-specific data can then lead to diminished service performance and harms to queer users.

Speech processing is one domain in which queer representation remains limited.[3] Processing spoken language is relevant for transgender and non-binary voices, since they challenge normative assumptions about gender (Zimman, 2018*, 2021*). Nevertheless, speech technology research focuses on binary genders and cisgender speech (Sanchez et al., 2024*). Voice also affects the expression and perception of sexual orientation (Levon, 2007*; Simpson and Weirich, 2020*).

Very few speech datasets explicitly contain speech by queer individuals, although such data might be included and simply not labeled. Although this hinders research on the robustness of speech tools, publicly sharing such metadata comes with privacy risks. Siegert et al. (2025)* present a German speech dataset of podcasts and videos by people who publicly identify as queer.[4] An audit of seven major audio corpora finds that queer identities are rarely discussed, and several datasets additionally contain queerphobic content (Agnew et al., 2025*).

Only a single project focuses on queer voices in the context of Automatic Speech Recognition: a recurring shared task on ASR for underrepresented social groups in India including transgender people (Bharathi et al., 2022). However, none of

[3]ACL affiliated conferences have been explicitly welcoming speech papers, yet none of the ACL Anthology venues appear to be the places where people choose to publish their works on queer speech tech. Subsequently, most of the papers covered in this section are not included in the 122 papers retrieved from the ACL Anthology, but can be found in the accompanying GitHub repository.

[4]Three more corpora without corresponding publications focus on transgender and/or non-binary English voices (German et al., 2022*; Dolquist and Munson, 2023*; Hope, 2023*). Another largely non-queer corpus contains metadata on gender expression and sexual orientation (Weirich et al., 2024*).

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

the overview or participant papers address gender, transness or queerness.

Three papers build speech synthesis systems for voices perceived as non-binary (Danielescu et al., 2023*; Szekely and Hope, 2024*; Hope and Székely, 2025*) using speech data from non-binary and trans speakers, and asking non-binary people to evaluate the resulting systems. Sigurgeirsson and Ungless (2024)* examine the shortcomings of voice cloning text-to-speech systems in replicating gay voices. They also discuss dual-use issues with queer speech technology, including their use for mockery.

Most publications on queer speech technology are very recent (2022–2025). The special session on Queer and trans speech science and technology[5] at Interspeech 2025 also demonstrates a current interest in analyzing and processing queer speech.[6] Given the (non-queer) research on ASR quality differences based on speaker gender (Feng et al., 2024*), one future research direction might be to investigate performance gaps for gender-non-normative voices, including in other tasks with speech input, such as spoken language understanding and speech translation. In speech translation or summarization, queer speakers might be especially prone to being misgendered, which has been acknowledged (Gaido et al., 2020), but not explicitly investigated.

Another domain in NLP where underrepresentation plays a pivotal role is research into the accessibility of information on queer topics via chatbots. Najafali et al. (2023)* investigate ChatGPT's ability to provide evidence-based recommendations for gender-affirming surgery according to the standards of care by the World Professional Association for Transgender Health. Although ChatGPT can accurately describe aspects of gender diversity and treatment for gender dysphoria, the study highlighted gaps in completeness, reliability, and explicit reference to authoritative guidelines. Koss et al. (2025)* present a similar user study, showing low trust in ChatGPT's ability to provide gender-affirming surgery information. Motivated by the high engagement of queer youth in social media, Brixey et al. (2017) develop a chatbot for HIV prevention and care for queer youth. They create a corpus of 3,000 question-and-answer pairs from reliable sources and train a chatbot which they deploy on Facebook Messenger. However, their work does not analyze their success in reaching queer youth.

Some works study chatbot use by queer individuals for mental health and emotional support. Lissak et al. (2024), Edwards et al. (2021), and Ma et al. (2024) conclude that these systems have positive effects, but are generic and culturally insensitive outside the US. While unsuitable to replace human emotional intelligence, the authors argue that a chatbot may be a useful tool for initial support.

Recommendations Our survey shows that speech technology and information access to queer topics are two domains where a lack of queer representation leads to potential harms for queer people. There is a tricky balance when it comes to inclusion and representation in these systems. In some cases inclusion can mean safety, e.g., when accurate queer health information is accessible via broad purpose chatbots. In other cases it could mean endangerment trough outing or the facilitation of surveillance, e.g. when speech recognition systems trained on queer voices disclose that a user is trans. Representation is not a panacea. We therefore recommend thinking of the consequences of inclusion of data or the collecting of gender and sexual orientation information, ideally in close collaboration with stakeholders.

4.2 Methods

4.2.1 Template-based Methods for Bias Detection

Large Language Models (LLMs), which increasingly serve as the backbone of modern NLP systems, are trained on massive, mostly uncurated online text corpora. LLMs thus tend to inherit and reproduce social biases from these corpora (Mickel et al., 2025)*. Examining LLMs and their outputs is therefor a common object of study in Queer NLP papers (30 papers, 24% of our survey).

Template-based methods are very prevalent in the examined papers, both for the measuring of bias and for the creation of training data. This can be seen as a product of the present time, as the affordances of LLMs make these methods expedient. As the most widely-used approach to elicit LLM bias, template-based methods involve prompting LLMs with sentences containing subject placeholders, which can be filled with terms representing

[5] sites.google.com/view/is-queer-trans/home

[6] Outside of NLP, Netzorg et al. (2024)* focused on how speaker identification systems struggle with voice modulation for gender expression and Doukhan et al. (2023)* and McAllister et al. (2025)* on systems that help with voice transition.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

various social groups, e.g., identity terms relating to gender and sexuality, names, or pronouns (Nozza et al., 2022; Dhingra et al., 2023; Hossain et al., 2023; Sosto and Barrón-Cedeño, 2024*; Ovalle et al., 2023a*). Notably, Felkner et al. (2023) take a community-centered template-based approach by collaborating with queer annotators to generate sentence pairs, such as "LGBTQ people are sinful" versus "Heterosexual people are sinful".

By analyzing the outputs conditioned on these different terms, the authors report systematic differences that often reflect social biases or imbalances. Given the complexity of assessing LLM bias in queer-related contexts, many studies rely on intrinsic bias measures. Building on template-based methods, common approaches include direct statistical comparison between expected and actual answers (Bunzeck and Zarrieß, 2024; Xie et al., 2024), as well as comparison of probabilistic scores such as pseudo-log-likelihood or perplexity for target-attribute pairings within sentence templates (Felkner et al., 2023; Hossain et al., 2023; Bergstrand and Gambäck, 2024; Xie et al., 2024; Ovalle et al., 2023a*).

In contrast, other studies perform group-level comparisons using downstream tasks, yielding more extrinsic bias measures. For instance, sentiment analysis has been widely applied to examine polarity (negative, neutral, or positive) in LLMs outputs (Kiritchenko and Mohammad, 2018; Sosto and Barrón-Cedeño, 2024*), while regard analysis (Sheng et al., 2019) has been used to capture perceived respectfulness in generated text (Ovalle et al., 2025; Dhingra et al., 2023). Similarly, harmful language detection tools such as Perspective API[7] are frequently used to detect toxicity and offensive language (Nozza et al., 2022; Sosto and Barrón-Cedeño, 2024*), alongside lexicon-based approaches such as HONEST (Nozza et al., 2021*), which builds on HurtLex (Bassignana et al., 2018*) to quantify offensiveness.

Several papers use other tasks as a testbed to assess the harmfulness of LLMs. For instance, Sobhani et al. (2023) evaluate gender bias in the context of hate speech and offensive language classification. In the same vein, some works also make the use of question-answering datasets for this purpose. SlayQA (Bunzeck and Zarrieß, 2024) provides contexts and question that specifically probe the model's ability to handle neopronouns, while BBQ (Parrish et al., 2022) provides contexts that are intentionally kept ambiguous as to expose LLMs' internal biases about different social groups. This is also looked into by Shaier et al. (2023), who analyze how the additional of superfluous demographic information changes the answer of LLM-based medical question-answering systems.

These methods are rarely tailored to capture the specific nuances of queer or intersectional identities in an usage context, which limits their ability to provide a full picture of harmful biases in LLMs.

Another under-examined problem is "unmarkedness" in language, as unnatural mentions of privileged identities in templates e.g., "straight man" may yield unreliable results (Blodgett et al., 2021). Here, a lack of engagement with linguistics may lead to insufficient methods.

Recommendations Introducing yet another dataset or template-based bias elicitation method offers limited additional insight. Contemporary LLMs can readily internalize such benchmarks to strengthen guardrails, which often makes the benchmark obsolete and results in sanitized yet still stereotypical outputs. What is needed instead are novel measurement paradigms that are grounded in real-world usage contexts and a stronger theoretical foundation for operationalizing phenomena such as erasure and heteronormativity, particularly once more overt manifestations of bias have been obscured.

4.2.2 Post-hoc Bias Mitigation

A central motivation in Queer NLP is to mitigate anti-queer biases in models. However, this is usually restricted to first revealing the existence of some bias, and only subsequently developing or applying methods that mitigate it. This reactive approach often results in a cat-and-mouse game, where papers demonstrate bias in existing models but are unable to develop better models, as debiasing methods are rarely adopted or feasible to operationalize at scale. This pattern has also been observed in other subfields of NLP, e.g., in interpretability research (Gururaja et al., 2023; Mosbach et al., 2024; Orgad et al., 2026).

The deficiencies of models can often be traced back to their training data, and better resources are thus one potential solution. For example, Ljubešić et al. (2020) introduce the LiLaH emotion lexicon for Croatian, Dutch, and Slovene, showing more consistent results on LGBTQ topics in emotion

[7] https://www.perspectiveapi.com.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

classification. Waldis et al. (2024) create a resource for German text classification for seven different tasks, including gender-neutral and gender-fair formulations.

In hate speech and toxicity detection, Engelmann et al. (2024) and Sahoo et al. (2022) construct datasets to detect dehumanizing language. Dacon et al. (2022) focus entirely on LGBTQIA+ individuals, creating a real-world dataset of harmful online conversational content. Wiegand and Ruppenhofer (2024) provide a dataset and train predictors for more subtle language where a group is portrayed as deviating from a norm. In a similar vein, Lu and Jurgens (2022) introduce a dataset to detect trans-exclusionary language by TERFs (see Glossary). Tang et al. (2024) create a debiasing dataset using Counterfactual Data Augmentation (CDA) to de-bias biased contexts. They construct anti-biased descriptors, i.e., replacing stereotypical associations with counter-stereotypical ones to generate texts with them, creating a dataset for fine-tuning.

On the side of de-biasing interventions that go beyond data, Dhingra et al. (2023)* apply a post-hoc method that combines chain-of-thought prompting with SHAP analysis (Lundberg and Lee, 2017*) to identify words in the LLM output that reduce social regard. After detecting low-regard words, they query an LLM for explanations and then re-prompt it to rephrase the sentence, removing those words while preserving meaning.

Recommendations Bias mitigation is an important part towards safe and fair NLP systems. However, rather than focusing solely on existing systems and their shortcomings, we recommend researchers to envision and co-create language technologies that are not merely harmless but actively affirming of queer identities and supportive to queer users. Grassroots efforts outside of academia exist in this regard, e.g., Shinigami Eyes[8], a browser extension that highlights transphobic and trans-friendly social network pages in different colors to help queer users navigate to safe online spaces. Bringing stakeholder ideas and use cases into research practice can prove as a fruitful way forward.

5 Current Gaps and Future Work

By characterizing common phenomena and methods, we also outline the gaps of existing research, some of which are mentioned in the cited papers as avenues for future work. In this section, we examine them and point to positive exemplars of first steps towards closing them, before moving on to more overarching topics.

5.1 Cis-normative conceptions of gender

An important thread of discussion in Queer NLP papers concerns the treatment of gender in research. Devinney et al. (2022) criticize the field's predominantly binary and cis-normative conceptions of gender. Hovy and Spruit (2016) note that researchers often use pronouns or names as a proxy for gender, and Gautam et al. (2024c) extend this observation to other sociodemographic attributes, describing ethical issues and giving practical advice. The framing of the treatment of gender data as an ethical issue has already been argued before by Larson (2017), outlining ethical principles about the use of gender, and suggesting that researchers make underlying theory of gender explicit and use gender only if necessary.

We would like to see these recommendations more widely adopted in NLP, especially when papers address gender bias. Gender bias dedicated venues such as the Workshop on Gender Bias in NLP[9] should release a set of guidelines for the treatment of gender as a variable (a "gender bias starter pack") to both guarantee a high standard within its accepted publications and to facilitate the permeation of best practices throughout the community.

5.2 Stakeholder Involvement

Although most papers name impact on the queer community as a motivating factor for their research, very few include queer community members as part of the development or evaluation of their methods[10]. More often, linguistic representations like identity terms, pronouns, and automatically-calculated metrics are used as a stand-in for users. Notable exceptions are: Gromann et al. (2023), who conduct a workshop with representatives from the intersex and non-binary communities to determine how they would like to be represented in translation; Felkner et al. (2023), who create the WinoQueer

[8] https://shinigami-eyes.github.io/

[9] https://gebnlp-workshop.github.io/

[10] While not a substitute for broad stakeholder engagement, Queer NLP research is often conducted by queer researchers drawing on lived experience. These researchers face a double bind: disclosing their positionality (see Appendix D) can strengthen the grounding of their work, but may also risk negative career consequences through being outed.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Benchmark with text input from the community; and Ungless et al. (2023), who survey trans and non-binary people on their experience with generated images, specifically asking for ways in which users would like to be represented. Nevertheless, these approaches engage with the stakeholders only once and do not establish ways in which the community is continuously involved in the stewardship of technologies.

There are two promising avenues through which Queer NLP can improve in the future. First, recent work on low-resource and endangered languages has increasingly emphasized stakeholder involvement (e.g., Bird and Yibarbuk, 2024), offering a useful model for Queer NLP. Second, relevant work on queer stakeholder involvement is likely also taking place in human–computer interaction venues, rather than in papers from the ACL Anthology, which primarily captures language technology interventions. Thus, future NLP research could benefit from drawing on findings from human–computer interaction literature.

5.3 Intersectionality and Interdisciplinarity

Ovalle et al. (2023b) point out that AI fairness literature mainly engages with intersectionality when applying metrics to demographic subgroups, and avoids other aspects, such as examining power structures. Similarly, Devinney (2025) argues that in order to mitigate the negative effects of stereotyping, NLP researchers need to engage with the social mechanisms behind it. However, we find that only 17% of all examined papers explicitly mention intersectionality.

We also find that the selection of studied languages limits the opportunity to engage with a variety of intersecting identities in the global queer community. Slightly less than half of the surveyed papers examine exclusively English language data, and even the inherently multilingual field of machine translation considers only translations to and from English (we examine non-English Queer NLP more closely in §5.4). Languages other than English skew towards Romance and Germanic languages (see Figure 3), which leads to a lack of cultural diversity and a concentration on concepts of queerness prevalent in the Global North.

Operationalizing intersectionality requires new data collection and processing practices and a deep engagement with the communities that they touch on, going beyond the currently practiced one-off involvement of stakeholders towards an integration of them in every step of the research process.

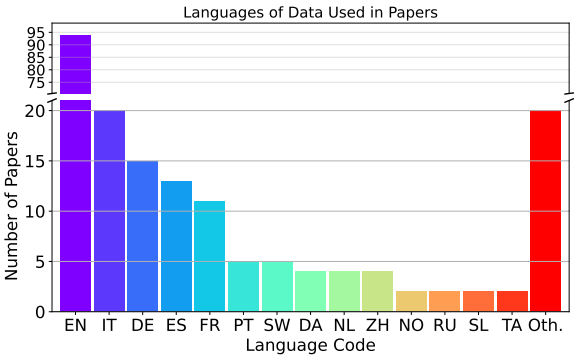


Figure 3: Breakdown of the 35 languages represented across papers. The 'Other' (Oth.) category aggregates all languages that appeared in only one paper. See Appendix B for a full list of languages along with paper counts.

NLP can be thought of as an inherently interdisciplinary field that straddles linguistics and computer science. Especially when dealing with queer language phenomena, previous work in sociolinguistics, gender studies, and queer studies can provide guiding principles for research. One positive example is Amironesei and Diaz (2023), who suggest modeling social relations as a pathway to improved hate speech detection.

While technological affordances shape much of what we see in our survey, we recommend researchers engage with the rich existing literature in the aforementioned fields to arrive at new ways of approaching problems and at better ways to connect social phenomena with data and model artifacts.

5.4 Non-English Queer NLP

An inherent limitation of our survey is the focus on ACL. Even with its international appeal, the kind of paper submitted and accepted is still geared towards an English-speaking audience (Lepp and Sarin, 2024), which leads to an unequal distribution of target languages (see Figure 3).

The predominance (if not hegemony) of English decreases when we broaden our scope beyond the Anglosphere. The Brazilian Computer Society has hosted workshops featuring papers on Queer NLP about and written in Portuguese (Souza et al., 2022; Santos et al., 2022), and so has the Italian Association of Computational Linguistics, although written in English (Nozza et al., 2023; Draetta et al., 2024). The Spanish Society for Natural Language Process-

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

ing is particularly prolific, with many Queer NLP papers on Spanish and other Iberian languages at their IberLEF workshop (Bel-Enguix et al., 2023; Jiménez-Zafra et al., 2023; Gómez-Adorno et al., 2024; García-Baena et al., 2024).

These workshops address topics that could be explored in English Queer NLP papers but are currently underrepresented. For example, García-Baena et al. (2023) introduce the task of hope speech detection for queer individuals in Spanish, defined as the detection of "speech that is able to relax a hostile environment and that helps, gives suggestions and inspires for good", which culminated in HOPE, a shared task on hope detection at IberLEF. The first edition in 2023 focused on queer hope speech (Jiménez-Zafra et al., 2023) and the second edition in 2024 still dedicated half of the task to it (García-Baena et al., 2024). While this topic is less explored in English Queer NLP, a notable exception is Pofcher et al. (2025). Similarly, the study of the identification of reclaimed queer slurs in Italian (Draetta et al., 2024) spawned a shared task on multilingual queer slur reclamation for the EVALITA workshop 2026, [11] a topic that remains understudied in English NLP (Zsisku et al., 2024).

Shared tasks play a key role in these events, drawing researchers' attention to the subject. The HOPE tasks from above included 8 papers in 2023 (Jiménez-Zafra et al., 2023) and 16 in 2024 (García-Baena et al., 2024), making it the largest repository of papers on hope detection for the queer community. There are similar numbers for hate detection in Mexican Spanish with the HOMO-MEX task at IberLEF (Bel-Enguix et al., 2023; Gómez-Adorno et al., 2024) and in Italian with the HODI task at EVALITA (Nozza et al., 2023). This may indicate that specific challenges help gather momentum for research in non-English Queer NLP, which can be especially useful for identities that don't fit the Western mold of "queer" e.g., hijra (see Glossary).

Much like English Queer NLP research influences the papers at venues not focused on English, many of their ideas and advances have relevance for languages beyond the one under study. We encourage researchers to broaden their focus accordingly.

5.5 Harms in the Wild vs. in Research

Most Queer NLP research has focused on a relatively narrow set of harms. Reliance on rigid taxonomies (e.g., of harm types, speech types, and identities) can flatten the queer experience while disregarding the fluidity of queer language and the necessity for expert stakeholder input. Creating these taxonomies is a fundamentally political decision that influences what is visible and invisible within a system (Bowker and Star, 2000; Smith et al., 2018; Guyan, 2025). Most hate speech and toxicity detection systems, for example, categorize (implicitly or explicitly) their targeted speech, but rarely involve the perspectives of queer users who experience hate speech: only 4 out of the 34 papers surveyed involve queer stakeholders.

Classification comes with potentially dire real-world consequences for queer people, given that queerness—as a deviation from what is constructed as "normal" or desirable—has historically been an important target of surveillance (Kafer and Grinberg, 2019), ostensibly on grounds of public health, reproduction, and morality (Conrad, 2009). Surveillance has an outsized effect on the most marginalized in our societies (Eubanks, 2018), and is further enhanced by several pursuits in NLP, including entity tracking, event tracking, sentiment analysis, and social media monitoring, all of which we have touched on in this survey. However, despite the way that NLP systems enable surveillance, there is little work that calls out these risks, let alone critiques our role as researchers in creating them (Kaffee et al., 2023).

In order to avoid the negative consequences of both (mis)classification and surveillance, queer users may reject the premise that systems should categorize them at all. Critical refusal, or opting out of being measured, labeled, or represented in normative ways, is a vital queer methodological position that current NLP systems leave little room for (Abramovich and Ma, 2024). Refusal is not simply a rejection of inclusion, but a rejection of the frameworks that inclusion demands: that identities be stable, categorizable, and intelligible to institutions (Costanza-Chock, 2020; Crawford, 2021).

In practice, this could include actions such as declining to specify a gender on a form, using intentionally ambiguous language online, resisting classification through irony, camp, or coded speech, or, among NLP researchers, refusing to use harmful categorization in research (Gautam et al., 2024c). These modes of resistance challenge NLP systems that depend on fixed typologies and clear labels, exposing a gap between what queer people want from

[11] https://multipride-evalita.github.io/

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

NLP technologies and what those technologies are designed to do. However, refusal is difficult to publish about in NLP, as it definitionally resists detection, annotation, and evaluation. As a result, queer refusal is rendered technically invisible even as it remains politically relevant.

5.6 Perspectives from Queer Theory

One aim of this survey is to look beyond immediate gaps in Queer NLP and to create visions for future work. To understand the complexity of socially-constructed identities and their interactions with sociotechnical systems, we can turn to queer theory approaches. This work examines and challenges power (Foucault, 1978) while confronting the certainty of sex, gender, and sexuality (Butler, 1990). Acknowledging that power plays a role in shaping what it means to be "queer" can help us better conceptualize why queer life is often denigrated or outright erased in NLP systems, providing a path forward for better Queer NLP research.

One of queer theory's aims is to question the stability of identity categories, taking particular issue with how these categories are presumed and normalized in everyday life and academic discourse. NLP is currently no less guilty of adhering to strict categorization, particularly when it comes to encoding queerness. In our survey, all papers that specifically study queerness rely on lists of sexual and gender identities, reducing nuanced or fuzzy forms of queer life into simple categorizations. Further, Klipphahn-Karge et al. (2023) argue that AI "posits societal norms within a double bind" by both *reflecting* existing social bias and *producing* heteronormative knowledge that flattens multiplicities. Because of this, some queer theorists claim that data science and its applications, such as NLP, are entirely incompatible with queerness (Keyes, 2019). Others, such as Shah (2023), propose ways to "queer" AI systems by reworking data science building blocks to allow imperfect, non-normative, and community-gathered data points. In either case, in order to break down cis- and heteronormative NLP standards, we must expand our scholarship and collaborations from the intersection of computing, statistics, and linguistics to include sociology, gender studies, queer theory, and beyond. With one foot "in the craft work of design" and the other "in the reflexive work of critique" (Agre, 1998), we can build interdisciplinary systems that proactively center those at the margins instead of relying on post-hoc tweaks and retroactive guardrails.

6 Conclusion

This work provides an overview of the current state of Queer NLP, summarizing findings from a variety of subtopics, as well as contextualizing them (§4–5). Limitations and our ethics statement can be seen in Appendix C and D.

Queer NLP echoes many shortcomings of the field as a whole. Despite some recent efforts (§5.4), slightly less than half work focuses exclusively on English, with most of the work centering English even when other languages are included (e.g., in machine translation).

Moreover, many works operate with a binary and cis-normative viewpoint of gender and sexuality, which could be informed better by research from queer theory. Finally, any harm mitigation usually happens post-hoc; this can at least partially be explained by the lack of stakeholder involvement in modern NLP systems.

For these reasons, we advocate for future work to radically rethink the ways in which we build language technologies. Going forward, researchers should develop solutions for socially-grounded rather than assumed understandings of harm and therefore involve affected queer communities. Community involvement needs to go beyond one-off workshops and data collection efforts towards an integration into all steps of the research pipeline, from the definition of use cases over model development to long-term ownership and auditing of resulting systems.

Our way forward can take inspiration from research on low-resource languages and human-computer interaction, through community-driven data collection and participatory action research, which involves stakeholders as collaborators and challenges oppressive epistemologies and methods. Moreover, interdisciplinary collaboration with researchers of sociology, gender studies, and queer theory can open new avenues to move beyond restrictive methods of data collection and model building. We call for the development of NLP systems that not only mitigate harm but actively support and uplift queer communities to make them the stewards of the language technologies that affect them.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Acknowledgments

We would like the reviewers and the action editor for their helpful comments. We would also like to thank all members of Queer in AI who supported us with advice throughout the whole process of writing this paper.

References

Shira Abramovich and Anna J. Ma. 2024. Evaluating refusal. Presented at the Workshop on Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI at NeurIPS 2024, Vancouver, Canada.

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H. Jiang, Ezra Awumey, and Sauvik Das. 2025. Sound check: Auditing recent audio dataset practices. In *Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES-25)*, pages 26–40.

Philip E. Agre. 1998. Toward a critical technical practice: Lessons learned in trying to reform AI. In Geoffrey Bowker, Susan Leigh Star, Les Gasser, and William Turner, editors, *Social Science, Technical Systems, and Cooperative Work*. Psychology Press.

Razvan Amironesei and Mark Diaz. 2023. Relationality and offensive speech: A research agenda. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 85–95, Toronto, Canada. Association for Computational Linguistics.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

Scott Andersen, Segio-Luis Ojeda-Trueba, Juan Vásquez, and Gemma Bel-Enguix. 2024. The Mexican gayze: A computational analysis of the attitudes towards the LGBT+ population in Mexico on social media across a decade. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 178–200, Mexico City, Mexico. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2253 of *CEUR Workshop Proceedings*, Torino, Italy. CEUR-WS.org.

Gemma Bel-Enguix, Helena Gómez-Adorno, Gerardo Sierra, Juan Vásquez, Scott Thomas Andersen, and Sergio Ojeda-Trueba. 2023. Overview of HOMO-MEX at Iberlef 2023: Hate speech detection in online messages directed towards the MEXican Spanish speaking LGBTQ+ population. *Procesamiento del Lenguaje Natural*, 71:361–370.

Selma Bergstrand and Björn Gambäck. 2024. Detecting and mitigating LGBTQIA+ bias in large Norwegian language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 351–364, Bangkok, Thailand. Association for Computational Linguistics.

B. Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N. Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 339–345, Dublin, Ireland. Association for Computational Linguistics.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Pa-*

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

pers), pages 1004–1015, Online. Association for Computational Linguistics.

Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT press.

Jacqueline Brixey, Rens Hoegen, Wei Lan, Joshua Rusow, Karan Singla, Xusen Yin, Ron Artstein, and Anton Leuski. 2017. SHIHbot: A Facebook chatbot for sexual health information on HIV/AIDS. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 370–373, Saarbrücken, Germany. Association for Computational Linguistics.

Bastian Bunzeck and Sina Zarrieß. 2024. The SlayQA benchmark of social reasoning: Testing gender-inclusive generalization with neopronouns. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 42–53, Miami, Florida, USA. Association for Computational Linguistics.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Thinking Gender. Routledge, New York.

Judith Butler. 2020. Critically queer. In Shane Phelan, editor, *Playing with Fire*, pages 11–29. Taylor & Francis.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Sky CH-Wang and David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9918–9938, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Rahul Ponnusamy, Malliga Subramanian, Paul Buitelaar, Miguel Ángel García-Cumbreras, Salud María Jimenez-Zafra, Jose Antonio Garcia-Diaz, Rafael Valencia-Garcia, and Nitesh Jindal. 2023. Overview of second shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 38–46, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality*. John Wiley & Sons.

Kathryn Conrad. 2009. 'Nothing to hide . . . nothing to fear': Discriminatory surveillance and queer visibility in Great Britain and Northern Ireland. In Noreen Giffney and Michael O'Rourke, editors, *The Ashgate Research Companion to Queer Theory*. Routledge.

Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press.

Kate Crawford. 2021. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(1).

Jamell Dacon, Harry Shomer, Shaylynn Crum-Dacon, and Jiliang Tang. 2022. Detecting harmful online conversational content towards LGBTQIA+ individuals. *CoRR*, cs.CL/2207.10032v1. Presented at the Queer in AI Workshop at NAACL 2022, Seattle, Washington, USA.

Andreea Danielescu, Sharone A Horowit-Hendler, Alexandria Pabst, Kenneth Michael Stewart, Eric M Gallo, and Matthew Peter Aylett. 2023. Creating inclusive voices for the 21st century: A non-binary text-to-speech for conversational assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany. Association for Computing Machinery.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hannah Devinney. 2025. Power(ful) associations: Rethinking "stereotype" for NLP. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 52–58, Vienna, Austria. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in NLP bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2083–2102, Seoul, Republic of Korea. Association for Computing Machinery.

Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *CoRR*, cs.CL/2307.00101v1. Presented at the Queer in AI Workshop at ACL 2023, Toronto, Canada.

Tanvi Dinkar, Aiqi Jiang, Simona Frenda, Poppy Gerrard-Abbott, Nancie A. Gunson, Gavin Abercrombie, and Ioannis Konstas. 2026. Can NLP tackle hate speech in the real world? Stakeholder-informed feedback and survey on counterspeech. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8276–8290, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devin V. Dolquist and Benjamin Munson. 2023. A palette of transmasculine voices. Data Repository for the University of Minnesota (DRUM).

David Doukhan, Simon Devauchelle, Lucile Girard-Monneron, Mía Chávez Ruz, V. Chaddouk, Isabelle Wagner, and Albert Rilliard. 2023. Voice passing: A non-binary voice gender prediction system for evaluating transgender voice transition. In *Proceedings of Interspeech 2023*, pages 5207–5211, Dublin, Ireland.

Lia Draetta, Chiara Ferrando, Marco Cuccarini, Liam James, and Viviana Patti. 2024. ReCLAIM project: Exploring Italian slurs reappropriation with large language models. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, volume 3878 of *CEUR Workshop Proceedings*, pages 335–342, Pisa, Italy. CEUR-WS.org.

Justin Edwards, Leigh Clark, and Allison Perrone. 2021. LGBTQ-AI? Exploring expressions of gender and sexual orientation in chatbots. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, Bilbao (online), Spain. Association for Computing Machinery.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Paul Engelmann, Peter Trolle, and Christian Hardmeier. 2024. A dataset for the detection of dehumanizing language. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 14–20, St. Julian's, Malta. Association for Computational Linguistics.

Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Rita Finkbeiner, Jörg Meibauer, and Heike Wiese. 2016. What is pejoration, and how can it be expressed in language? In Rita Finkbeiner, Jörg Meibauer, and Heike Wiese, editors, *Pejoration*, pages 1–18. John Benjamins Publishing Company.

Michel Foucault. 1978. *The History of Sexuality*, volume 1: An Introduction. Pantheon Books, New York.

Steinunn Rut Friðriksdóttir. 2024. The GenderQueer test suite. In *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340, Miami, Florida, USA. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Daniel García-Baena, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, José Antonio García-Díaz, and Rafael Valencia-García. 2023. Hope speech detection in Spanish: The LGBT case. *Language Resources and Evaluation*, 57:1487–1514.

Daniel García-Baena, Fazlourrahman Balouchzahi, Sabur Butt, Miguel Ángel García-Cumbreras, Atnafu Lambebo Tonja, José Antonio García-Díaz, Selen Bozkurt, Bharathi Raja Chakravarthi, Hector G. Ceballos, Rafael Valencia-García, Grigori Sidorov, L. Alfonso Ureña-López, Alexander Gelbukh, and Salud María Jiménez-Zafra. 2024. Overview of HOPE at IberLEF 2024: Approaching hope speech detection in social media from two perspectives, for equality, diversity and inclusion and as expectations. *Procesamiento del Lenguaje Natural*, 73:407–419.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024a. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 2024b. WinoPron: Revisiting English Winogender schemas for consistency, coverage, and grammatical case. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 52–66, Miami, Florida, USA. Association for Computational Linguistics.

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. 2024c. Stop! In the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

James Sneed German, Maria Candea, LeAnn Brown, Timothy Mahrt, and Oriana Reid-Collins. 2022. Gender spectrum speech corpus. ORTOLANG (Open Resources and TOols for LANGuage).

Ina Goel. 2016. Hijra communities of Delhi. *Sexualities*, 19(5-6):535–546.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <MASK>: Evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Dagmar Gromann, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, Igor Miladinovic, Sigrid Schefer-Wenzl, Daniela Duh, and Katharina Bühn. 2023. Participatory research as a path to community-informed, gender-fair machine translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 49–59, Tampere, Finland. European Association for Machine Translation.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.

Kevin Guyan. 2025. *Rainbow Trap: Queer Lives, Classifications and the Dangers of Inclusion*. Bloomsbury Academic Publishing.

Helena Gómez-Adorno, Gemma Bel-Enguix, Hiram Calvo, Sergio Ojeda-Trueba, Scott Thomas Andersen, Juan Vásquez, Tania Alcántara, Miguel Soto, and Cesar Macias. 2024. Overview of HOMO-MEX at IberLEF 2024: Hate speech detection towards the Mexican Spanish speaking LGBT+ population. *Procesamiento del Lenguaje Natural*, 73:393–405.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

Amanda Hicks, Michael Rutherford, Christiane Fellbaum, and Jiang Bian. 2016. An analysis of WordNet's coverage of gender identity using Twitter and the national transgender discrimination survey. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 123–130, Bucharest, Romania. Global Wordnet Association.

Jacob Hobbs. 2025. Theories of "sexuality" in natural language processing bias research. Unpublished B.A. thesis, Department of Computer Science, University of Virginia.

Maxwell Hope. 2023. Mid-Atlantic gender expansive speech (MAGES) corpus. Online.

Maxwell Hope and Éva Székely. 2025. Voices of 'cyborg awesomeness': Posthuman embodiment of nonbinary gender expression in AI speech technologies. In *Proceedings of Interspeech 2025*, pages 689–693, Rotterdam, The Netherlands.

Adnan Hossain. 2017. The paradox of recognition: Hijra, third gender and sexual rights in Bangladesh. *Culture, Health & Sexuality*, 19(12):1418–1431.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Human Rights Campaign. 2025. Glossary of terms. Retrieved October 28th, 2025.

Salud María Jiménez-Zafra, Miguel Ángel Garcia-Cumbreras, Daniel García-Baena, José Antonio Garcia-Díaz, Bharathi Raja Chakravarthi, Rafael Valencia-García, and Luis Alfonso Ureña-López. 2023. Overview of HOPE at IberLEF 2023:

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Multilingual hope speech detection. *Procesamiento del Lenguaje Natural*, 71:371–381.

Rosemary A. Joyce. 2015. Heteronormativity. In *The International Encyclopedia of Human Sexuality*, pages 501–581. John Wiley & Sons.

Gary Kafer and Daniel Grinberg. 2019. Editorial: Queer surveillance. *Surveillance & Society*, 17(5):592–601.

Lucie-Aimée Kaffee, Arnav Arora, Zeerak Talat, and Isabelle Augenstein. 2023. Thorny roses: Investigating the dual use dilemma in natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13977–13998, Singapore. Association for Computational Linguistics.

Os Keyes. 2019. Counting the countless. Published in the online magazine *Real Life*. Retrieved October 28th, 2025.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Klipphahn-Karge, Ann-Kathrin Koster, and Sara Morais dos Santos Bruss. 2023. Introduction: Queer AI. In *Queer Reflections on AI*, pages 1–19. Routledge.

Mitchell R. Koss, Matthew McLaughlin, Kayla Switalla, Israel Falade, and Esther Kim. 2025. Exploring the role of ChatGPT in decision making for gender-affirming surgery. *Artificial Intelligence Surgery*, 5(3):116–125.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411, Torino, Italia. European Language Resources Association and International Committee on Computational Linguistics.

Manuel Lardelli and Dagmar Gromann. 2023. Gender-fair post-editing: A case study beyond the binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland. European Association for Machine Translation.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? How commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Haley Lepp and Parth Sarin. 2024. A global AI community requires language-diverse publishing. *CoRR*, cs.CL/2408.14772v2. Presented at the Global AI Cultures workshop at ICLR 2024, Vienna, Austria.

Erez Levon. 2007. Sexuality in context: Variation and the sociolinguistic perception of identity. *Language in Society*, 36(4):533–554.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2040–2079, Mexico City, Mexico. Association for Computational Linguistics.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The LiLaH emotion lexicon of Croatian, Dutch and Slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 153–157, Barcelona, Spain (Online). Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A cross-lingual study of homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Christina Lu and David Jurgens. 2022. The subtle language of exclusion: Identifying the toxic speech of trans-exclusionary radical feminists. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 79–91, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 4768–4777, Long Beach, California, USA. Curran Associates Inc.

Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z. Gajos. 2024. Evaluating the experience of LGBTQ+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, Hawaii, USA. Association for Computing Machinery.

Tara McAllister, Collin Eagen, Yi Shan, Peter Traver, Daphna Harel, Tae Hong Park, and Vesna Novak. 2025. Web-based application for real-time biofeedback of vocal resonance in gender-affirming voice training: Design and usability evaluation. In *Proceedings of Interspeech 2025*, pages 664–668, Rotterdam, The Netherlands.

Jennifer Mickel, Maria De-Arteaga, Liu Leqi, and Kevin Tian. 2025. More of the same: Persistent representational harms under increased representation. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 61277–61315. Curran Associates, Inc.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva. 2024. From insights to actions: The impact of interpretability and analysis research on NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3105, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Najafali, Chandler Hinson, Justin M. Camacho, Logan G. Galbraith, Tannon L. Tople, Danielle Eble, Brielle Weinstein, Loren S. Schechter, Amir H. Dorafshar, and Shane D. Morrison. 2023. Artificial intelligence knowledge of evidence-based recommendations in gender affirmation surgery and gender identity: Is ChatGPT aware of WPATH recommendations? *European Journal of Plastic Surgery*, 46(6):1169–1176.

Robin Netzorg, Alyssa Cote, Sumi Koshin, Klo Vivienne Garoute, and Gopala Krishna Anumanchipalli. 2024. Speech after gender: A trans-feminine perspective on next steps for speech science and technology. In *Proceedings of Interspeech 2024*, pages 3075–3079, Kos, Greece.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Nonbinary Wiki. 2025a. Cissexism. Retrieved October 28th, 2025.

Nonbinary Wiki. 2025b. LGBT. Retrieved October 28th, 2025.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, Anne Lauscher, and Dirk Hovy. 2022. Measuring harmful sentence completion in language models for LGBTQIA+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 26–34, Dublin, Ireland. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. HODI at EVALITA 2023: Overview of the first shared task on homotransphobia detection in Italian. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, volume 3473 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Hadas Orgad, Fazl Barez, Tal Haklay, Isabelle Lee, Marius Mosbach, Anja Reusch, Naomi Saphra, Byron C. Wallace, Sarah Wiegreffe, Eric Wong, Ian Tenney, and Mor Geva. 2026. Position: Interpretability can be actionable. In *Forty-third International Conference on Machine Learning Position Paper Track*.

Organizers of Queer in AI, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, and 31 others. 2023. Queer in AI: A case study in community-led participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pages 1882–1895, Chicago, Illinois, USA. Association for Computing Machinery.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023a. "I'm fully who I am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1246–1266, Chicago, Illinois, USA. Association for Computing Machinery.

Anaelia Ovalle, Krunoslav Lehman Pavasovic, Louis Martin, Luke Zettlemoyer, Eric Michael Smith, Kai-Wei Chang, Adina Williams, and Levent Sagun. 2025. The root shapes the fruit: On the persistence of gender-exclusive harms in aligned language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 3094–3105, New York, NY, USA. Association for Computing Machinery.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023b. Factoring the matrix of domination: A critical review and reimagination of intersectionality in AI fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 496–511, Montréal, Canada. Association for Computing Machinery.

Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland. European Association for Machine Translation.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. An LLM-as-a-judge approach for scalable gender-neutral translation evaluation. In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 46–63, Geneva, Switzerland. European Association for Machine Translation.

Jonathan Pofcher, Christopher M. Homan, Randall Sell, and Ashiqur R. KhudaBukhsh. 2025. Hope vs. hate: Understanding user interactions with LGBTQ+ news content in mainstream US news media through the lens of hope speech. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19862–19888, Suzhou, China. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

A Pranav, Alissa Valentine, Shaily Bhatt, Yanan Long, Arjun Subramonian, Amanda Bertsch, Anne Lauscher, and Ankush Gupta, editors. 2025. *Proceedings of the Queer in AI Workshop*. Association for Computational Linguistics, Hybrid format (in-person and virtual).

Krithika Ramesh, Sumeet Kumar, and Ashiqur Khudabukhsh. 2022. Revisiting queer minorities in lexicons. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 245–251, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. GATE: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pages 845–854, Montréal, Canada. Association for Computing Machinery.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ariadna Sanchez, Alice Ross, and Nina Markl. 2024. Beyond the binary: Limitations and possibilities of gender-related speech technology research. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 526–532.

Vinícius Santos, Felipe Henriques, and Gustavo Guedes. 2022. O discurso de ódio homofóbico no Twitter a partir da análise de dados. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining (BRASNAM)*, pages 109–120, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação (SBC).

Allison Scott, Freada Kapor Klein, and Uriridiakoghene Onovakpuri. 2017. Tech leavers study: A first-of-its-kind analysis of why people voluntarily left jobs in tech. Technical report, Kapor Center for Social Impact.

Nishant Shah. 2023. I spy, with my little AI: How queer bodies are made dirty for digital technologies to claim cleanness. In *Queer Reflections on AI*, pages 57–62. Routledge.

Sagi Shaier, Kevin Bennett, Lawrence Hunter, and Katharina Kann. 2023. Emerging challenges in personalized medicine: Assessing demographic effects on biomedical question answering systems. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 540–550, Nusa Dua, Bali. Association for Computational Linguistics.

Deepawali Sharma, Vedika Gupta, and Vivek Kumar Singh. 2023. Detection of homophobia & transphobia in Malayalam and Tamil: Exploring deep learning methods. In *Advanced Network Technologies and Intelligent Computing: Part II*, volume 1798 of *Communications in Computer and Information Science (CCIS)*, pages 217–226, Varanasi, India. Springer Nature Switzerland.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Ingo Siegert, Jan Marquenie, and Sven Grawunder. 2025. Queer Waves: a German speech dataset capturing gender and sexual diversity from podcasts and YouTube. In *Proceedings of Interspeech 2025*, pages 679–683, Rotterdam, The Netherlands.

Atli Sigurgeirsson and Eddie L. Ungless. 2024. Just because we camp, doesn't mean we should: The ethics of modelling queer voices. In *Proceedings of Interspeech 2024*, pages 3050–3054, Kos, Greece.

Adrian P. Simpson and Melanie Weirich. 2020. Phonetic correlates of sex, gender and sexual orientation. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Warren Smith, Matthew Higgins, George Kokkinidis, and Martin Parker. 2018. Becoming invisible: The ethics and politics of imperceptibility. *Culture and Organization*, 24(1):54–73.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Nasim Sobhani, Kinshuk Sengupta, and Sarah Jane Delany. 2023. Measuring gender bias in natural language processing: Incorporating gender-neutral linguistic forms for non-binary gender identities in abusive speech detection. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1121–1131, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mae Sosto and Alberto Barrón-Cedeño. 2024. QueerBench: Quantifying discrimination in language models toward queer identities. *CoRR*, cs.CL/2406.12399v1. Presented at the Queer in AI Workshop at NAACL 2024, Mexico City, Mexico.

Andrey Souza, Eduardo Nakamura, and Fabíola Nakamura. 2022. Detecção de discurso de ódio: Homofobia. In *Anais do XVI Brazilian e-Science Workshop*, pages 73–80, Porto Alegre, RS, Brasil. Sociedade Brasileira de Computação (SBC).

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *CoRR*, cs.CL/2112.14168v1.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? Assessing bias against same-gender relationships in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 365–375, Bangkok, Thailand. Association for Computational Linguistics.

Arjun Subramonian, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat. 2024. Understanding "democratization" in NLP and ML research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3151–3166, Miami, Florida, USA. Association for Computational Linguistics.

Eva Szekely and Maxwell Hope. 2024. An inclusive approach to creating a palette of synthetic voices for gender diversity. In *Proceedings of Interspeech 2024*, pages 3070–3074, Kos, Greece.

Kunsheng Tang, Wenbo Zhou, Jie Zhang, Aishan Liu, Gelei Deng, Shuai Li, Peigui Qi, Weiming Zhang, Tianwei Zhang, and NengHai Yu. 2024. GenderCARE: A comprehensive framework for assessing and reducing gender bias in large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, CCS '24, pages 1196–1210, Salt Lake City, Utah, USA. Association for Computing Machinery.

Jordan Taylor, Ellen Simpson, Anh-Ton Tran, Jed R. Brubaker, Sarah E Fox, and Haiyi Zhu. 2024. Cruising queer HCI on the DL: A literature review of LGBTQ+ people in HCI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, Hawaii, USA. Association for Computing Machinery.

Joshua Tint. 2025. Guardrails, not guidance: Understanding responses to LGBTQ+ language in large language models. In *Proceedings of the Queer in AI Workshop at NAACL 2025*, pages 6–16, Albuquerque, New Mexico, USA (Hybrid). Association for Computational Linguistics.

Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Juan Vásquez, Scott Andersen, Gemma Bel-enguix, Helena Gómez-adorno, and Sergio-luis Ojeda-trueba. 2023. HOMO-MEX: A Mexican Spanish annotated corpus for LGBT+phobia detection on Twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A rewriting approach for gender inclusivity in Portuguese. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore. Association for Computational Linguistics.

Andreas Waldis, Joel Birrer, Anne Lauscher, and Iryna Gurevych. 2024. The Lou dataset - exploring the impact of gender-fair language in German text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10604–10624, Miami, Florida, USA. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

Angelina Wang. 2025. Identities are not interchangeable: The problem of overgeneralization in fair machine learning. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 485–497, Athens, Greece. Association for Computing Machinery.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 336–349, Seoul, Republic of Korea. Association for Computing Machinery.

Melanie Weirich, Daniel Duran, and Stefanie Jannedy. 2024. Gender and age based f0-variation in the German Plapper corpus. In *Proceedings of Interspeech 2024*, pages 1565–1569, Kos, Greece.

Michael Wiegand and Josef Ruppenhofer. 2024. Oddballs and misfits: Detecting implicit abuse in which identity groups are depicted as deviating from the norm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2200–2218, Miami, Florida, USA. Association for Computational Linguistics.

Hsiao-Han Wu and Shu-Kai Hsieh. 2017. Exploring lavender tongue from social media texts in Chinese. In *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017)*, pages 68–80, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing.

Sean Xie, Saeed Hassanpour, and Soroush Vosoughi. 2024. Addressing healthcare-related racial and LGBTQ+ biases in pretrained language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4451–4464, Mexico City, Mexico. Association for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Amy (Azure) Zhou. 2024. Queer bias in natural language processing: Towards more expansive frameworks of gender and sexuality in NLP bias research. *GRACE: Global Review of AI Community Ethics*, 2(1).

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

Lal Zimman. 2018. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, 12(8):e12284.

Lal Zimman. 2021. Gender diversity and the voice. In Jo Angouri and Judith Baxter, editors, *The Routledge Handbook of Language, Gender, and Sexuality*, pages 69–90. Routledge.

Eszter Zsisku, Arkaitz Zubiaga, and Haim Dubossarsky. 2024. Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, pages 241–249, Stuttgart, Germany. Association for Computing Machinery.

Nadezhda Zueva, Madina Kabirova, and Pavel Kalaidin. 2020. Reducing unintended identity bias in Russian hate speech detection. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 65–69, Online. Association for Computational Linguistics.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

A Glossary

Here, we define some terms that we use throughout the paper:

- **cisnormativity**: an ideology that promotes cisgender individuals as valid while painting transgender individuals as "unnatural or dangerous" (Nonbinary Wiki, 2025a)
- **heteronormativity**: an ideology that promotes only heterosexual individuals as natural and normal (Joyce, 2015)
- **hijra**: a third gender of "male-bodied feminine-identified people" (Hossain, 2017) rooted in the social, cultural and religious framework of India and several other South Asian countries (Goel, 2016)
- **LGBTQIA+**: a collective term for gender, sex and sexuality minorities. It is an acronym for **L**esbians, **G**ay, **B**isexual, **T**ransgender, **Q**ueer/**Q**uestioning, **I**ntersexual and **A**romatic/**A**sexual, with the plus sign (**+**) including any other relevant minority identity not hereby covered (Human Rights Campaign, 2025; Nonbinary Wiki, 2025b)
- **non-binary**: a category of individuals who do not fully identify with a binary gender i.e., individuals who are neither man or woman
- **trans-exclusionary radical feminists (TERFS)**: individuals who advocate for "feminism" with the view that transgender women are infringing on the rights of cisgender women

B Language Breakdown

This is a breakdown of all the languages covered in the examined papers:

Language of Data	Count	Singletons	
English	94	code-mixed Urdu	Farsi
Italian	20	Korean	code-mixed English-Tamil
German	15	Telugu	Icelandic
Spanish	13	Kannada	Greek
French	11	Roman Urdu	Filipino
Portuguese	5	Maltese	Czech
Swedish	5	Indian English	Croatian
Chinese	4	Hindi	Turkish
Dutch	4	Catalan	Slovak
Danish	4	Polish	Gujarati
Slovene	2	Russian	
Tamil	2	Norwegian	
Norwegian	2		

Table 1: Counts of appearances of languages across papers. Languages with only one paper are grouped under "Singletons."

C Limitations

In our efforts to survey the field of NLP for works regarding queer topics, we acknowledge two main limitations: Firstly, while we include papers from other venues such as Interspeech, FAccT, CHI, etc. in Section 4 and non-English conferences in Section 5.4, our work is still centered on research published in *ACL conferences. Secondly, our paper search utilized a list of queer identity keywords. Both factors might lead to a selection that misses publications from other venues, other non-English venues, and queer aspects that are not captured by our chosen keywords.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026

D Ethics Statement

Positionality The authors of this paper hold many different standpoints afforded to us by our diverse identities, cultures, geographic locations, and personal histories. Most relevantly for this work, we draw from our positions both as experts in various NLP fields and as members and allies of queer communities. Our perspective on the issues we identify and discuss does not aspire to be an objective "view from nowhere" (Haraway, 1988), but instead is rooted in our local knowledge and understandings. In particular, this paper is largely shaped by North American and Eurocentric perspectives. We draw together partial perspectives to synthesize a collective sense of the current state of Queer NLP, its implications and impacts, and possible futures.

E Keywords

We use the ACL Anthology Python library to retrieve all papers whose titles or abstracts contain at least one of the keywords below (case-insensitive matching with no lemmatization or word tokenization): `"agender"`, `"aromantic"`, `"bisexual"`, `"gay"`, `"gender-ambiguous"`, `"gender fair"`, `"gender-fair"`, `"gender inclusiv"`, `"gender-inclusive"`, `"gender-neutral"`, `"gender non"`, `"genderqueer"`, `"glbt"`, `"heteronormative"`, `"homophobia"`, `"homotransphob"`, `"homotransphobia"`, `"lesbian"`, `"lgbt"`, `"non-binary"`, `"nonbinary"`, `"queer"`, `"same-gender"`, `"trans-exclusionary"`, `"transgender"`, `"transphobia"`, `"transphobic"`.

Downloaded from direct.mit.edu/tacl/article-pdf/doi/10.1162/TACL.a.818/2637753/tacl.a.818.pdf by guest on 09 November 2026